

Université de Montréal

Alternative Strategies for Deciphering the Genetic Architecture of Childhood
Pre-B Acute Lymphoblastic Leukemia

par

Jasmine Healy

Programmes de biologie moléculaire
Faculté de médecine

Thèse présentée à la Faculté de médecine
en vue de l'obtention du grade de doctorat
en biologie moléculaire

Juin, 2010

© Jasmine Healy, 2010

Université de Montréal
Faculté de médecine

Cette thèse intitulée :

Alternative Strategies for Deciphering the Genetic Architecture of Childhood
Pre-B Acute Lymphoblastic Leukemia

présentée par:

Jasmine Healy

a été évaluée par un jury composé des personnes suivantes:

Dr. Bernard Brais

.....
président-rapporteur

Dr. Daniel Sinnett

.....
directeur de recherche

Dr. Guillaume Lettre

.....
membre du jury

Dr. David Goldgar

.....
examineur externe

Dr. Patrick Cossette

.....
représentant du doyen

FRENCH SUMMARY

La leucémie lymphoblastique aigüe (LLA) est une maladie génétique complexe. Malgré que cette maladie hématologique soit le cancer pédiatrique le plus fréquent, ses causes demeurent inconnues. Des études antérieures ont démontrées que le risque à la LLA chez l'enfant pourrait être influencé par des gènes agissant dans le métabolisme des xénobiotiques, dans le maintien de l'intégrité génomique et dans la réponse au stress oxydatif, ainsi que par des facteurs environnementaux. Au cours de mes études doctorales, j'ai tenté de disséquer davantage les bases génétiques de la LLA de l'enfant en postulant que la susceptibilité à cette maladie serait modulée, au moins en partie, par des variants génétiques agissant dans deux voies biologiques fondamentales : le point de contrôle G1/S du cycle cellulaire et la réparation des cassures double-brin de l'ADN. En utilisant une approche unique reposant sur l'analyse d'une cohorte cas-contrôles jumelée à une cohorte de trios enfants-parents, j'ai effectué une étude d'association de type gènes/voies biologiques candidats. Ainsi, j'ai évalué le rôle de variants provenant de la séquence promotrice de 12 gènes du cycle cellulaire et de 7 gènes de la voie de réparation de l'ADN, dans la susceptibilité à la LLA. De tels polymorphismes dans la région promotrice (pSNPs) pourraient perturber la liaison de facteurs de transcription et mener à des différences dans les niveaux d'expression des gènes pouvant influencer le risque à la maladie.

En combinant différentes méthodes analytiques, j'ai évalué le rôle de différents mécanismes génétiques dans le développement de la LLA chez l'enfant. J'ai tout d'abord étudié les associations avec gènes/variants indépendants, et des essais fonctionnels ont été effectués afin d'évaluer l'impact des pSNPs sur la liaison de facteurs de transcription et l'activité promotrice allèle-spécifique. Ces analyses ont mené à quatre publications. Il est peu probable que ces gènes de

susceptibilité agissent seuls; j'ai donc utilisé une approche intégrative afin d'explorer la possibilité que plusieurs variants d'une même voie biologique ou de voies connexes puissent moduler le risque de la maladie; ces travaux ont été soumis pour publication. En outre, le développement précoce de la LLA, voir même *in utero*, suggère que les parents, et plus particulièrement la mère, pourraient jouer un rôle important dans le développement de cette maladie chez l'enfant. Dans une étude par simulations, j'ai évalué la performance des méthodes d'analyse existantes de détecter des effets fœto-maternels sous un design hybride trios/cas-contrôles. J'ai également investigué l'impact des effets génétiques agissant via la mère sur la susceptibilité à la LLA. Cette étude, récemment publiée, fût la première à démontrer que le risque de la leucémie chez l'enfant peut être modulé par le génotype de sa mère.

En conclusions, mes études doctorales ont permis d'identifier des nouveaux gènes de susceptibilité pour la LLA pédiatrique et de mettre en évidence le rôle du cycle cellulaire et de la voie de la réparation de l'ADN dans la leucémogénèse. À terme, ces travaux permettront de mieux comprendre les bases génétiques de la LLA, et conduiront au développement d'outils cliniques qui amélioreront la détection, le diagnostic et le traitement de la leucémie chez l'enfant.

Mots clés : leucémie lymphoblastique aigüe de l'enfant, épidémiologie génétique, susceptibilité génétique, polymorphisme régulateur, expression génique, cycle cellulaire, réparation de l'ADN, voie biologique, interaction gène-gène, association génétique fœto-maternelle

ENGLISH SUMMARY

Childhood acute lymphoblastic leukemia (ALL) is a complex and heterogeneous genetic disease. Although it is the most common pediatric cancer, its etiology remains poorly understood. Previous studies provided evidence that childhood ALL might originate through the collective contribution of different genes controlling the efficiency of carcinogen metabolism, the capacity of maintaining DNA integrity and the response to oxidative stress, as well as environmental factors. In my doctoral research project I attempted to further dissect the genetic intricacies underlying childhood ALL. I postulated that a child's susceptibility to ALL may be influenced, in part, by functional sequence variation in genes encoding components of two core biologic pathways: G1/S cell cycle control and DNA double-strand break repair. Using a unique two-tiered study design consisting of both unrelated ALL cases and healthy controls, as well as case-parent trios, I performed a pathway-based candidate-gene association study to investigate the role of sequence variants in the promoter regions of 12 candidate cell cycle genes and 7 DNA repair genes, in modulating ALL risk among children. Polymorphisms in promoter regions (pSNPs) could perturb transcription factor binding and lead to differences in gene expression levels that in turn could modify the risk of disease.

To better depict the complex genetic architecture of childhood ALL, I used multiple analytical approaches. First, individual genes/variants were tested for association with disease, while functional *in vitro* validation was performed to evaluate the impact of the pSNPs on differential transcription factor binding and allele-specific promoter activity. These analyses led to four published articles. Given that these genes are not likely to act alone to confer disease risk I used an integrative approach to explore the possibility that combinations of functionally relevant pSNPs among several components of the same or of

interconnected pathways, could contribute to modified childhood ALL risk either through pathway-specific or epistatic effects; this work was recently submitted for publication. Finally, childhood ALL is thought to arise *in utero* suggesting that the parents, and in particular the mother, may play an important role in shaping disease susceptibility in their offspring. Using simulations, I investigated the performance of existing methods to test for maternal genotype associations using a case-parent trio/case-control hybrid design, and then assessed the impact of maternally-mediated genetic effects on ALL susceptibility among children. This published work was the first to show that the mother's genotype can indeed influence the risk of leukemia in children, further corroborating the importance of considering parentally-mediated effects in the study of early-onset diseases.

In conclusion, my doctoral work lead to the identification of novel genetic susceptibility loci for childhood ALL and provided evidence for the implication of the cell cycle control and DNA repair pathways in leukemogenesis. Better elucidation of the genetic mechanisms underlying the pathogenesis of ALL in children could be of great diagnostic value and provide data to help guide risk-directed therapy and improve disease management and outcome. Ultimately, this study brings us one step closer to unraveling the genetic architecture of childhood ALL and provides a stepping-stone towards disease prevention.

Key words: childhood acute lymphoblastic leukemia, genetic epidemiology, genetic susceptibility, promoter SNP, gene expression, cell cycle, DNA repair, pathway, gene-gene interaction, maternally-mediated genotype effect

TABLE OF CONTENTS

LIST OF FIGURES	xi
LIST OF TABLES	xiii
LIST OF ABBREVIATIONS.....	xvii
DEDICATION.....	xix
PREFACE – Childhood Acute Lymphoblastic Leukemia in the Age of Genomics	xx
CHAPTER ONE - Introduction.....	1
Childhood Cancer	2
Leukemia	3
The Natural History of Childhood Acute Lymphoblastic Leukemia	7
Incidence, Survival and Trends	7
Pathobiology of Childhood Acute Lymphoblastic Leukemia	8
Molecular Genetic Alterations	9
Prenatal Origin	12
Cooperating Oncogenic Lesions	12
Risk Factors	14
Environmental Exposures	16
Genetic Susceptibility Factors	23
Dissecting the Genetic Basis of Complex Diseases	27
The Landscape of Genetic Variation	28
Genetic Association Studies.....	35
Case-control Association Studies.....	37
Linkage Disequilibrium	38
Family-Based Association Studies	42
Hypothesis-Driven versus Hypothesis-Free Approaches	43
Finding the Missing Heritability	45
Epistasis.....	46
Maternal Genetic Effects.....	47
Candidate Cancer Pathways.....	48
Objectives of The Project.....	57
Expected Impact of the Project.....	58
References	59
CHAPTER TWO – Individual Main Effects of Candidate Genes on the Susceptibility of Childhood Acute Lymphoblastic Leukemia	77
Candidate Gene Association Studies	78
Genome-Wide Association Studies	90

ARTICLE I - Promoter SNPs in G1/S Checkpoint Regulators and their Impact on the Susceptibility to Childhood Leukemia	92
Author Contributions.....	93
Abstract	94
Introduction.....	95
Patients, Materials, and Methods	96
Results	99
Discussion	103
Figures	107
Tables.....	110
References	118
Reprint Permissions	124
ARTICLE II - No Evidence for Association Between <i>TGFB1</i> Promoter SNPs and the Risk of Childhood Pre-B Acute Lymphoblastic Leukemia Among French Canadians	125
Author Contributions.....	126
Letter to the Editor	127
Tables.....	130
References	132
Reprint Permissions	134
ARTICLE III - Functional Impact of Sequence Variation in the Promoter Region of <i>TGFB1</i>	135
Author Contributions.....	136
Abstract	137
Introduction.....	138
Material and Methods	139
Results	142
Discussion	146
Figures	151
Tables.....	155
References	157
Reprint Permissions	160
ARTICLE IV - Replication Analysis Confirms the Association of <i>ARID5B</i> with Childhood B-Cell Acute Lymphoblastic Leukemia	163
Author Contributions.....	164
Abstract	165
Introduction.....	166
Design and Methods	167
Results and Discussion	169
Tables.....	174
References	182
Reprint Permissions	184

CHAPTER THREE – Combined Pathway Effects and Gene-Gene Interactions in the Susceptibility to Childhood Acute Lymphoblastic Leukemia	185
ARTICLE V - Promoter Variants in Genes Involved in the Cell Cycle and DNA Repair Pathways and the Susceptibility to Childhood Acute Lymphoblastic Leukemia	188
Author Contributions	189
Abstract	190
Introduction.....	191
Material and Methods	192
Results	201
Discussion	203
Tables.....	209
References	226
Reprint Permissions	230
CHAPTER FOUR – The Role of Maternally-Mediated Genetic Effects in the Susceptibility to Childhood Acute Lymphoblastic Leukemia	231
ARTICLE VI - Detection of Fetomaternal Genotype Associations in Early-Onset Disorders: Evaluation of Different Methods and their Application to Childhood Leukemia	234
Author Contributions.....	235
Abstract	236
Introduction.....	237
Materials and Methods	240
Results and Discussion	245
Conclusions.....	254
Figures	256
Tables.....	265
References	275
Reprint Permissions	279
CHAPTER FIVE – Conclusions and Future Perspectives.....	280
Summary and Discussion of Main Findings.....	281
A Candidate Pathway Strategy.....	291
Parental Genetics in Early-Onset Disorders	295
Limitations of the Study	298
Future Perspectives.....	300
Integrative Genomics	300
Gene-environment interactions	301
Candidate Gene Versus Genome-Wide Association Studies	302
Common Versus Rare Variants.....	303
Conclusion.....	306
References	308

APPENDIX I – Measures of Association	xxiv
APPENDIX II – Power of the Association Studies and Complete Results of the Case-Control and Family-Based Association Tests.....	xxxii
APPENDIX III Results of the Fetomaternal Association Tests for the DNA Double-Strand Break Repair Genes	lxv
APPENDIX IV Follow-Up Analysis of a Genome-Wide Association Study Identifies CDKN2A as a Susceptibility Locus for Childhood Acute Lymphoblastic Leukemia	lxxi
CURRICULUM VITAE	lxxviii

LIST OF FIGURES

CHAPTER ONE

Figure 1. Hematopoietic differentiation	4
Figure 2. Chromosomal translocations in childhood ALL	11
Figure 3. Greaves' multi-stage model for the development of childhood ALL ...	13
Figure 4. Critical windows of exposure for childhood ALL.....	17
Figure 5. Single nucleotide polymorphism	29
Figure 6. Gene-based single nucleotide polymorphisms	30
Figure 7. The impact of a SNP in a transcription factor binding site (TFBS)	32
Figure 8. Rationale for rSNP discovery	33
Figure 9. Disease mapping strategies.....	36
Figure 10. Rationale for association studies	40
Figure 11. Cellular responses to environmental exposures and cancer susceptibility.....	49
Figure 12. Illustration of the G1/S cell cycle checkpoint.....	51
Figure 13. DNA repair mechanisms	55

CHAPTER TWO

Figure 1. Design of the candidate gene association study	78
Figure 2. Candidate genes selected based on their function in two cancer- related pathways.....	84
Figure 3. Power calculations for main effects using a case-control design	88
Figure 4. Power calculations for main effects using a family-based case-parent trio design	89

ARTICLE I

Figure 1. Polymorphisms detected in the promoter regions of <i>CDKN2A</i> , <i>CDKN2B</i> , <i>CDKN1A</i> , and <i>CDKN1B</i>	107
Figure 2. EMSA illustrating allelic DNA-protein interactions in the promoter region of <i>CDKN2B</i>	108

ARTICLE III

Figure 1. <i>TGFB1</i> haplotype network	151
Figure 2. Gene reporter assays evaluating promoter activity of 4 promoter haplotypes of <i>TGFB1</i>	152
Figure 3. Representative EMSA analyses showing allelic DNA-protein interactions for pSNPs -1886G>A, -1571A>G, -1550DEL/AGG and -509C>T in Jeg-3.....	153

CHAPTER FOUR

ARTICLE VI

Figure 1. Cohort data used for simulations	256
Figure 2. Type I error rates and power for the maternal association test under mating asymmetry.....	257
Figure 3. Power of the forward stepwise procedure to detect joint fetomaternal associations	258
Figure 4. Type I error rate for the HD and HD-NPC approaches for the maternal association test assuming either mating symmetry or asymmetry	259
Figure 5. Power of HD and HD-NPC to detect mating asymmetry.....	260
Figure 6. Log-linear, likelihood-ratio association analysis between 29 regulatory SNPs from 12 cell cycle genes and childhood pre-B acute lymphoblastic leukemia (ALL).....	262
Supplementary Figure 1. Relative distribution of reciprocal mating types as a function of mating asymmetry	263
Supplementary Figure 2. Type I error rates for fetomaternal association testing under mating asymmetry	264

CHAPTER FIVE

Figure 1. Details of the linkage disequilibrium region at 9p21.3.....	285
Figure 2. The V(D)J recombination mechanism.....	289
Figure 3. Cellular responses to environmental exposures and childhood ALL susceptibility.....	294
Figure 4. A paradigm shift in genetic association testing – from common to rare variants	305

APPENDIX II

Figure I. Allelic associations of cell cycle checkpoint gene variants with childhood ALL	xl
Figure II. Allelic associations of DNA double-strand break repair gene variants with childhood ALL.....	xlvi

APPENDIX III

Figure I. Log-linear, likelihood-ratio association analysis between 20 regulatory SNPs from 7 DNA double-strand break repair genes and childhood pre-B acute lymphoblastic leukemia	lxvi
---	------

LIST OF TABLES

CHAPTER ONE

Table 1. Childhood leukemia – a heterogeneous group of disorders categorized by cell type and level of differentiation	5
Table 2. Main chromosomal changes in childhood acute lymphoblastic leukemia	9
Table 3. Exposure-dependent risk factors reported to be significantly associated with childhood ALL.....	19
Table 4. Genetic risk factors reported to be significantly associated with childhood ALL	25

CHAPTER TWO

Table 1. Characteristics of the pre-B ALL patients from the Quebec Childhood ALL cohort	80
Table 2. Characteristics of genes and corresponding DNA variants genotyped in the association studies.....	85

ARTICLE I

Table 1. Characteristics of the PCR primers and ASO probes used to genotype the selected pSNPs	110
Table 2. Allele frequencies of promoter SNPs in <i>CDKN2A</i> , <i>CDKN2B</i> , <i>CDKN1A</i> , and <i>CDKN1B</i> in childhood pre-B ALL patients and controls	111
Table 3. Distribution of <i>CDKN2A</i> , <i>CDKN2B</i> , <i>CDKN1A</i> , and <i>CDKN1B</i> promoter-based genotypes among childhood patients with pre-B ALL and controls	112
Table 4. Impact of promoter SNPs on predicted transcription factor binding sites	113
Table 5. Distribution of <i>CDKN2B</i> , <i>CDKN1A</i> , and <i>CDKN1B</i> promoter haplotypes in patients with pre-B ALL and controls.....	114
Table 6. Family-based association analysis of promoter SNPs in <i>CDKN2A</i> , <i>CDKN2B</i> , <i>CDKN1A</i> , and <i>CDKN1B</i>	116
Table 7. FBAT analysis of promoter haplotypes in <i>CDKN2B</i> , <i>CDKN1A</i> , and <i>CDKN1B</i>	117

ARTICLE II

Table 1. Allele and genotype frequencies of pSNPs in <i>TGFB1</i> in childhood pre-B ALL patients and controls.	130
Table 2. Distribution of <i>TGFB1</i> promoter haplotypes in pre-B ALL patients and controls.	131

ARTICLE III

Table 1. <i>TGFB1</i> promoter haplotype distributions among five continental population groups	155
---	-----

Table 2. Impact of pSNPs on predicted transcription factor binding sites and summary of EMSA results	156
--	-----

ARTICLE IV

Table 1. Characteristics of the B-cell ALL patients from the Quebec Childhood ALL cohort	174
--	-----

Table 2. Replication analysis in the Quebec Childhood ALL cohort of germline SNPs whose allele frequencies differed between children with ALL and control groups in two genome-wide association studies	175
---	-----

Table 3. Distribution of <i>ARID5B</i> genotypes among B-cell ALL cases and controls from the Quebec Childhood ALL cohort and gender-specific genotype risks estimates	177
--	-----

Supplementary Table 1. Summary of primers used in the PCR and allele-specific primer extension (ASPE) assays for SNP genotyping	179
---	-----

Supplementary Table 2. Distribution of <i>ARID5B</i> haplotypes among B-cell ALL cases and controls from the Quebec Childhood ALL cohort and gender-specific haplotype risks estimates.....	180
---	-----

CHAPTER THREE

ARTICLE V

Table 1. Gens and DNA variants genotyped in B-cell ALL patients, their parents, and healthy controls.....	209
---	-----

Table 2. Prior matrix used in the hierarchical model	211
--	-----

Table 3. Characteristics of the French-Canadian pre-B ALL patients from the Quebec Childhood ALL cohort.....	213
--	-----

Table 4. Main effects of pSNPs on pre-B ALL risk among children, as estimated by conventional maximum-likelihood analysis and hierarchical modeling	214
---	-----

Supplementary Table 1. Distribution of DNA repair and cell cycle genotypes among B-cell ALL cases and controls from the Quebec Childhood ALL cohort and their effect on ALL risk, as estimated by logistic regression.....	216
--	-----

Supplementary Table 2. Main effects of genetic polymorphisms on ALL risk among children, as estimated by hierarchical modeling with a range of second-stage residual variance values.....	220
---	-----

Supplementary Table 3. Gene-gene interaction effects of selected polymorphisms on ALL risk among children, as estimated by hierarchical modeling with a range of second-stage residual variance	222
---	-----

Supplementary Table 4. Gene-gene interaction effects on pre-B ALL risk among children, as estimated by conventional maximum-likelihood analysis and hierarchical modeling	224
---	-----

CHAPTER FOUR

ARTICLE VI

Table 1. Forward stepwise likelihood-ratio testing procedure used to dissect child and maternal genotype associations	265
---	-----

Table 2. The eight simulation models used for evaluation of the fetomaternal association tests	266
Table 3. Parental genotype distributions under mating symmetry and mating asymmetry	267
Table 4. Genes and DNA variants genotyped in the pre-B ALL association study	268
Table 5. Distribution of <i>CDKN2A</i> rs36228834 and <i>CDKN2B</i> rs36229158 genotypes and associated risk estimates for pre-B ALL susceptibility among children	270
Supplementary Table 1. Log-linear, likelihood-ratio association analysis between 29 regulatory SNPs from 12 cell cycle genes and childhood pre-B ALL	272

APPENDIX I

Table I. Presentation of data from a cohort study or a case-control study of a binary risk factor.....	xxv
Table II. Presentation of data from a family-based case-trio study of a binary risk factor.....	xxix

APPENDIX II

Table I. Power calculations for main effects using a case-control design	xxxiii
Table II. Power calculations for main effects in the family-based study design	xxxv
Table III. Power calculations for gene-gene interaction effects in the case-control study design	xxxviii
Table IV. Allele frequencies of promoter SNPs in G1/S cell cycle checkpoint genes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk	xli
Table V. Distribution of G1/S cell cycle genotypes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk, as estimated by logistic regression	xlili
Table VI. Distribution of G1/S cell cycle haplotypes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk...	xlvi
Table VII. Allele frequencies of promoter SNPs in DNA double-strand break repair genes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk	xliv
Table VIII. Distribution of DNA double-strand break repair genotypes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk, as estimated by logistic regression.....	li
Table IX. Distribution of DNA double-strand break repair haplotypes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk	liv
Table X. Family-based association analysis of promoter SNPs in G1/S cell cycle checkpoint genes	lvii

Table XI. Family-based association analysis of promoter haplotypes in G1/S cell cycle checkpoint genes.....	lix
Table XII. Family-based association analysis of promoter SNPs in DNA double-strand break repair genes.....	lxi
Table XIII. Family-based association analysis of promoter haplotypes in DNA double-strand break repair genes	lxii

APPENDIX III

Table I. Log-linear, likelihood-ratio association analysis between 20 regulatory SNPs from 7 double-strand break repair genes and childhood pre-B ALL.....	lxvii
--	-------

LIST OF ABBREVIATIONS

ALG, automatic luminex genotyping
ALL, acute lymphoblastic leukemia
AML, acute myelogenous leukemia
ASO, allele-specific oligonucleotide hybridization
ASPE, allele-specific primer extension
BER, base excision repair
bp, basepair
CDCV, common disease common variant hypothesis
CDKI, cyclin-dependent kinase inhibitor
CEPG, conditioning on exchangeable parental genotypes
CI, confidence interval
CLP, common lymphoid progenitor
CLR, conditional logistic regression
CML, chronic myelogenous leukemia
CMP, common myeloid progenitor
CNV, copy-number variant
CRM, *cis*-regulatory module
cSNP, coding single nucleotide polymorphism
df, degree of freedom
dHPLC, denaturing high-performance liquid chromatography
DNA, deoxyribonucleic acid
DSBR, double-strand break repair
EM, expectation-maximization
EMSA, electrophoretic mobility shift assay
FAB, French-American-British classification system
FBAT, family-based association test
FDR, false discovery rate
GxG, gene-gene interaction
GMP, granulocytic myelomonocytic progenitor
GRR, genotype relative risk
GWAS, genome-wide association study
HD, hybrid design
HM, hierarchical modeling
HR, homologous recombination
HSC, hematopoietic stem cell
HWE, Hardy-Weinberg equilibrium
IALLGC, International Acute Lymphoblastic Leukemia Genetics Consortium

IBD, identity by descent
ICCC, International Classification of Childhood Cancer
indel, insertion/deletion
iSNP, intronic single nucleotide polymorphism
kb, kilobase
LD, linkage disequilibrium
LEM, Log-linear and event history analysis with missing data using the EM algorithm
LT-HSC, long-term hematopoietic stem cell
LR, likelihood ratio
MA, mating asymmetry
MAF, minor allele frequency
MEP, megakaryocytic/erythroid progenitor
MH, Mantel-Haenszel
MMR, mismatch repair
MPP, multipotent progenitor cell
MRV, multiple rare variant hypothesis
MS, mating symmetry
NER, nucleotide excision repair
NHEJ, non-homologous end-joining
NK, natural killer
OR, odds ratio
PCR, polymerase chain reaction
PGL, pathway genetic load
pHap, promoter haplotype
pSNP, promoter single nucleotide polymorphism
QcALL, Quebec Childhood Acute Lymphoblastic Leukemia Cohort
RAF, risk allele frequency
RNA, ribonucleic acid
RR, relative risk
rSNP, regulatory single nucleotide polymorphism
SD, standard deviation
SNP, single nucleotide polymorphisms
ST-HSC, short-term hematopoietic stem cell
TF, transcription factor
TFBS, transcription factor binding site
UTR, untranslated region

*I dedicate this thesis to my parents, Deborah and Patrick,
whose love and support nurtured and guided me along this journey;
and to my two best friends, my sister Rose and my brother Zachary,
for holding my hand every step of the way.*

PREFACE

*CHILDHOOD ACUTE LYMPHOBLASTIC
LEUKEMIA IN THE AGE OF GENOMICS*

“Cancer research driven by the allure of miracle cures is impoverished if it does not pay equal attention to possible causal mechanisms and prospects for prevention.”

— M. Greaves, 1999.

Over the past decade, a key focus of cancer research has been geared toward dissecting variation in cancer predisposition through the identification of inherited genetic changes that influence cancer risk, with the ultimate goal of decreasing mortality by reducing risk and improving diagnosis and treatment. Following the sequencing of the human genome, we moved rapidly into the age of genomics leading us from linkage analysis that were successful for identifying high-risk gene mutations involved in familial cancers to ever-growing association studies that now allow us to identify low to moderate risk alleles involved in sporadic cancers. From genotyping a handful of variants in a few candidate genes, to genotyping millions of variants genome-wide, to sequencing entire genomes, rapid technological advances have allowed us to peer ever so deeper into the genomes of cancer cells and into the genomes of individuals genetically predisposed to cancer, shedding light on the complex underpinnings of this multifaceted disease.

Childhood acute lymphoblastic leukemia (ALL) is one of the great success stories of modern medicine; thirty years ago a child that was diagnosed with ALL had little chances of survival, today through the application of intensive multiagent chemotherapeutic regimens, the cure rate for childhood ALL exceeds an impressive 80% in the developed world. This achievement stems mainly from advances in our knowledge of the pathobiology of the disease. Several clinical and biological features that correlate with outcome are now used to assign risk-

based treatment regimes to ALL sufferers. However a significant proportion of patients still fail therapy for unknown reasons and the long-term effects of the intense chemotherapeutic cocktails that are administered to patients can in certain instances be as debilitating as the disease itself. And on the downside of this success story is the shadowed fact that, in developed countries, ALL is still the leading cause of death by disease among children. More striking still is the fact that despite success in treating the disease, very little is known of the etiology of childhood ALL; and to better treat the disease is to better understand it.

ALL results from a series of mutational events within an immature blood cell that halt cell maturation and eventually lead to malignant proliferation and disruption of normal blood production. Over 200 genetic alterations have been identified so far in ALL tumors, with a handful of recurrent chromosomal rearrangements and mutations characterizing most of the cases. While there is well-established evidence that these mutational events play an important role in driving the leukemic process, the events leading up to leukemogenesis are not known. Initiation of the disease occurs during fetal life or in early infancy and, as with most other cancers, is likely caused by a combination of environmental and genetic factors. The assertion that ALL has a genetic basis has long been pursued through association studies based on candidate genes. These studies were recently complemented by genome-wide studies that vindicated the role of common inherited genetic variation in childhood ALL susceptibility.

There is no doubt that the genomics era will have a profound impact on the diagnosis and therapy of childhood ALL. But with progress comes new challenges; already the issue of missing heritability is daunting researchers interested in dissected the genetic architecture of complex diseases. In the research presented here I explore the effects of genetic variation on childhood

ALL susceptibility by addressing the following questions. Can we deviate from traditional analytical approaches to further illuminate the genetic basis of childhood ALL? Can deregulation of core cellular functions such as cell division and maintenance of genomic integrity influence a child's susceptibility to ALL? How important are the mother's genes in shaping her offspring's susceptibility to disease? And in light of the growing popularity of agnostic genome-wide searches, can candidate gene approaches still help explain some of the interindividual variability in complex disease susceptibility? Using a unique design that involves collecting DNA from childhood ALL cases as well as their parents and unrelated control individuals, I investigate the role of genes involved in two cancer-related pathways, the cell cycle control and DNA repair mechanisms, in childhood ALL. This pathway-based candidate gene association study provides a unique opportunity to investigate some of the genetic subtleties involved in this pediatric disorder. It is my hope that this research will provide greater insight into the etiologic intricacies of childhood ALL and help pave the way toward new opportunities for prevention. Unraveling the genetic architecture of childhood ALL will bring us one step closer to that ultimate goal of decreasing mortality by reducing risk and improving diagnosis and treatment.

CHAPTER ONE

INTRODUCTION

CHILDHOOD CANCER

Cancer is thought of as a disease of ageing, one in which the DNA of a normal cell accumulates sufficient mutations that it acquires a selective advantage and becomes capable of uncontrolled, unlimited proliferation. Yet cancer is also the leading cause of death by disease among children (1). It is estimated that each year, approximately 150 children out of every million children younger than 20 years of age will be diagnosed with cancer (2). And while incidence rates have been increasing steadily since the mid-1970's, the etiology of many childhood cancers remains elusive.

Cancer in children differs markedly from its adult counterpart with regard to cancer type, site of occurrence, as well as clinical behavior. Beyond the shorter latency period observed in pediatric cancers, they are often more aggressive and more invasive. The majority of tumors diagnosed in children stem from immature "embryonic-like" cell types, whereas adult cancers are mainly carcinomas that arise in epithelial tissue (3). Consequently, a more appropriate classification system was developed for childhood cancers based on cell morphology and tissue of origin rather than on the primary anatomical site of appearance as in adult cancers (4). The most common pediatric neoplasms are leukemias (cancer of the blood) representing 25% of all cancer cases among children younger than 20 years of age, followed by brain and central nervous system cancers (17%) and lymphomas (cancer of the lymphatic system) (16%) (5). The relative contribution of leukemia to the total childhood cancer burden rises as high as 46% among children aged 2-3 years, making this disease the leading cause of cancer-related deaths in children.

LEUKEMIA

Hematopoiesis is the highly regulated and hierarchical process during which blood cells are formed. Self-renewing progenitors in the bone marrow, the hematopoietic stem cells (HSCs), give rise to multipotent progenitors that in turn produce lineage-committed progenitors that give rise to the mature blood cells of either the myeloid or lymphoid lineages (Figure 1). In leukemia, normal hematopoiesis is disrupted. Development of normal hematopoietic cells is arrested in an early stage of differentiation in the bone marrow, and malignant proliferation of the immature lymphoid or myeloid cells depletes the pool of functionally mature blood cells and eventually invades the blood, lymph nodes, central nervous system and other vital organs (6). Leukemia is a clonal disease arising from the neoplastic transformation of a single cell and in many respects the initial leukemia cell resembles a stem cell with unlimited proliferation potential and self renewal capabilities and the ability to give rise to a new, albeit abnormal, hematopoietic tissue (7).

Leukemia is a heterogeneous group of neoplasms classified on the basis of cell type of origin. Leukemias can arise during any step of the hematopoietic process and are either acute, aggressive diseases affecting mostly immature, undifferentiated cells, or chronic, less rapidly progressing diseases affecting more mature and differentiated hematopoietic cells. The three major classifications of childhood leukemia are acute lymphoblastic leukemia (ALL), accounting for 75% to 80% of childhood leukemia cases; acute myelogenous leukemia (AML), accounting for 20% to 25% of cases; and chronic myelogenous leukemia (CML), accounting for ~3% of cases (6) (Table 1).

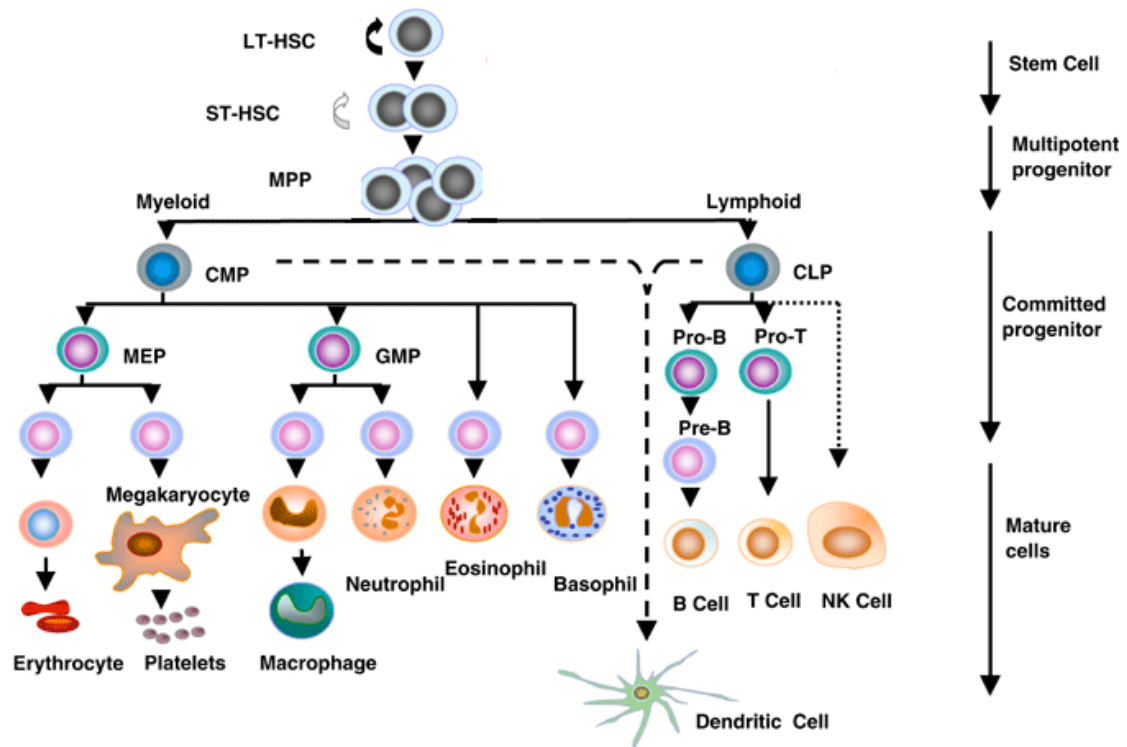


Figure 1. Hematopoietic differentiation

As HSCs divide, they can either self-renew, commit to differentiation or die by apoptosis in order to maintain a steady-state level of HSCs in the bone marrow and constantly provide progenitors for the various haematological lineages. The HSC population is made up of long-term HSCs (LT-HSCs) capable of lifetime self-renewal and short-term HSCs (ST-HSCs) that only briefly self-renew and give rise to multipotent progenitor cells (MPPs) which differentiate into the mutually exclusive myeloid and lymphoid lineages. The common myeloid progenitors (CMPs) give rise to the myelomonocytic progenitors (GMPs) which produce macrophages and granulocytes, and to the megakaryocytic/erythroid progenitors (MEPs) which produce megakaryocytes, platelets and erythrocytes. The common lymphoid progenitors (CLPs) give rise to B and T lymphocytes and natural killer (NK) cells. Both CMPs and CLPs can give rise to dendritic cells. Cell surface markers can be used to discriminate between all stem and progenitor hematopoietic cell populations. (7)

Figure adapted from Larsson & Karlsson, 2005 (8).

Table 1. Childhood leukemia – a heterogeneous group of disorders categorized by cell type and level of differentiation

Diagnostic Group	Specific Diagnosis	Rate per million ^a	% of Cases
Lymphoid leukemia		29.2	100.0
	Lymphoid leukemia, NOS		0.2
	Acute lymphoblastic		99.2
	Subacute lymphoid		0.0
	Chronic lymphocytic		0.1
	Aleukemic lymphoid		0.2
	Prolymphocytic leukemia		0.1
	Burkitt's cell leukemia		0.5
	Lymphosarcoma		0.0
Acute non-lymphocytic		7.6	100.0
	Erythroleukemia		0.4
	Acute erythremia		0.2
	Acute myeloid leukemia		68.7
	Aleukemic myeloid		0.0
	Acute promyelocytic		7.1
	Acute myelomonocytic		9.3
	Acute monocytic		9.1
	Aleukemic monocytic		0.0
	Acute megakaryoblastic		5.1
Chronic myeloid leukemia		1.3	100.0
	Chronic myeloid leukemia		98.6
	Chronic myelomonocytic		1.4
Other specified leukemias		0.2	100.0
	Plasma		0.0
	Chronic erythremia		0.0
	Myeloid leukemia, NOS		33.3
	Subacute myeloid		0.0
	Basophilic leukemia		0.0
	Eosinophilic leukemia		0.0
	Monocytic leukemia, NOS		8.3
	Subacute monocytic		0.0
	Chronic monocytic		0.0
	Mast cell leukemia		0.0
	Myeloid sarcoma		58.3
	Acute panmyelosis		0.0
	Acute myelofibrosis		0.0
	Hairy cell leukemia		0.0
Unspecified leukemias		1.2	100.0
	Leukemia, NOS		20.5
	Acute leukemia, NOS		79.5
	Subacute leukemia, NOS		0.0
	Chronic leukemia, NOS		0.0
	Aleukemia leukemia, NOS		0.0

Diagnostic groups and subcategories are based on the International Classification of Childhood Cancer (ICCC) definitions. Incidence rates adjusted

to the 1970 US standard population. In bold are the three major leukemia subtypes diagnosed in children. NOS, not otherwise specified.

^a Age-adjusted incidence rates and percent distribution for specific diagnostic subcategories of leukemia based on data from the Surveillance Epidemiology and End Results, 1975-95 for patients age <20.

Table adapted from Smith et al., 1999 (5).

THE NATURAL HISTORY OF CHILDHOOD ACUTE LYMPHOBLASTIC LEUKEMIA

Acute lymphoblastic leukemia (ALL) is the most frequent pediatric cancer and is itself a diverse group of diseases distinct both biologically and clinically. ALL can arise in either the B or T lineage of lymphocyte cells (Figure 1). Classification of B- and T-cell ALL is done based on cell morphology, using cytological features distinguished by the French-American-British (FAB) classification system (9) – including cell size, nuclear chromatin, nuclear shape, nucleoli, amount of cytoplasm, basophilia of cytoplasm, and cytoplasmic vacuolation – and using immunophenotyping of lineage- and maturation-specific cell surface antigens via flow cytometry (10). B-lineage ALL can be further subdivided into pro-B cell ALL (also known as early pre-B and pre-pre-B ALL) and pre-B ALL (also known as cALL for common ALL or simply as B-cell ALL) (11). Pro-B ALL is one of the most immature ALL subtypes and occurs mainly in infants aged birth to 1 year; it is rare and accounts for only ~5% of childhood ALL cases (12). B-cell ALL is the most common subtype, accounting for 80%, while T-cell ALL represents approximately 15% of all newly diagnosed pediatric ALL cases (13).

Incidence, Survival and Trends

It is estimated that approximately 1,100 children are diagnosed with ALL each year in Canada (14). While ALL occurs in children worldwide, its incidence varies between nations with Costa Rica, Finland and Canada experiencing the highest rates, China and India the lowest (15). As for most childhood cancers, ALL is slightly more prevalent in boys compared to girls with a 1.2:1 male to female ratio, except for pro-B ALL which exhibits a slight predominance in females (12). Incidence peaks between 2-5 years of age and tends to be higher in socioeconomically developed populations compared to developing countries

that tend to have lower rates with no obvious age-specific incidence peak (16). A two- to three-fold higher incidence rate is observed in white children compared to black children and rates among Hispanics are highest of all, suggesting differences in disease frequency associated with race and/or socioeconomic status (17).

Childhood ALL is one of the great cancer success stories of the 20th century. While less than 20% of diagnosed childhood ALL patients survived their disease 40 years ago, modern treatment protocols have managed to completely reverse survival rates. Among patients receiving contemporary chemotherapy treatment, the five-year survival rate is now 80% (18). However, 20% of the cases remain resilient to current treatment protocols and ultimately succumb to their disease. These numbers also conceal that, of the patients that become five-year survivors of childhood ALL, a substantial number develop long-term treatment-related complications including death (19, 20). Therefore, while the marked improvement in the overall cure rate for ALL is undoubtedly very impressive, treatment is far from being optimal. And importantly, despite decreasing mortality rates, the incidence of ALL among children younger than 20 years of age has been increasing with improved socio-economic conditions (5).

Pathobiology of Childhood Acute Lymphoblastic Leukemia

Success in treatment is due in large part to our increased understanding of the pathobiology of ALL. Leukemia is a disease of the genome characterized by gross genomic and chromosomal alterations which ultimately provide the leukemic cell with a selective and proliferative advantage.

Molecular Genetic Alterations

Although the primary causes of childhood ALL remain unknown, the mechanisms through which the disease arises can be postulated based on specific genetic alterations found in ALL. Over 200 somatic genetic alterations have been identified in ALL but for the most part they are characterized by gross chromosomal changes such as changes in DNA ploidy, chromosomal translocations and deletions. The acquired genetic lesions lead to precise stage-specific developmental arrest and allow unlimited self-renewal and clonal expansion of the B or T progenitor cell. Table 2 summarizes the main molecular alterations observed in ALL. The most frequent are hyperdiploidy (>50 chromosomes) and the t(12;21)/*ETV6-AML1* gene fusion; together they constitute 50% of all childhood ALL and almost 80% of ALL occurring between the ages of 2-5 years (21-23).

Table 2. Main chromosomal changes in childhood acute lymphoblastic leukemia

Cell type	Chromosome abnormality	Molecular lesion	Frequency (%)	Functional product
Pro-B (in infants)	11q23 translocations	<i>MLL-AF4</i> , <i>MLL-ENL</i> and other fusions	85 (of infant ALL)	Modified transcription factor
B-cell	Hyperdiploidy	Increased gene dosage	35 (of B-cell ALL)	Unknown
	t(12;21)	<i>ETV6-AML1</i> fusion	20 (of B-cell ALL)	Chimeric transcription factor
	t(1;19)	<i>E2A-PBX1</i> fusion	5 (of B-cell ALL)	Chimeric transcription factor
	t(9;22)	<i>BCR-ABL</i> fusion	5 (of B-cell ALL)	Activated kinase
T-cell	1q deletion; t(1;14)	<i>SIL-SCL</i> fusion	25 (of T-cell ALL)	Dysregulated transcription factor

Table adapted from Greaves & Wiemels, 2003 (24).

Hyperdiploid ALL is characterized by the nonrandom gain of chromosomes X, 4, 6, 10, 14, 17, 18, and 21. Though hyperdiploidy is suspected to incur a selective advantage to leukemic cells through gene dosage effects, the precise mechanism through which it occurs and its impact on leukemogenesis are unknown (23).

Chromosomal translocations are initiated by DNA double-strand breaks that occur simultaneously in a single cell. Through the process of repair, fusion gene products are formed resulting in functionally viable chimeric proteins with altered or dysregulated function (24). Several recurrent translocations are found in childhood ALL patients (Figure 2). The most prevalent is the t(12;21) translocation which creates a fusion gene that involves the transcriptional repressor *ETV6* and the hematopoietic-specific transcription factor *AML1*. The resulting ETV6-AML1 chimera leads to impaired hematopoietic differentiation, however the precise mechanisms by which it causes leukemia remain elusive.

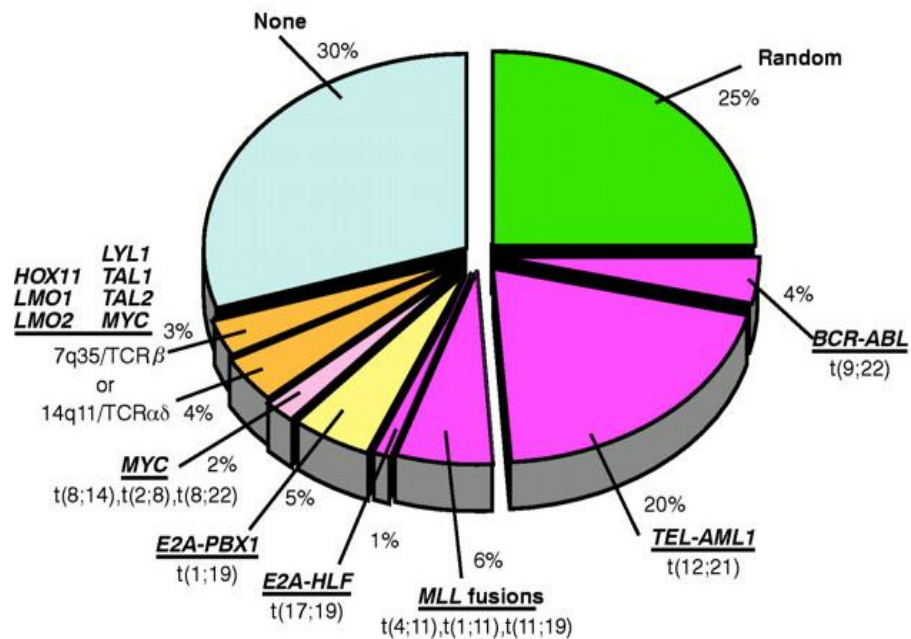


Figure 2. Chromosomal translocations in childhood ALL

The t(12;21) translocation which gives rise to the *TEL/AML1* (also known as the *ETV6/AML1*) fusion gene is the most prevalent among pre-B ALL cases. Note however that while chromosomal translocations are frequent in leukemia, roughly 30% of childhood ALL cases harbor no apparent alterations.

Figure adapted from Pui et al., 2004 (25).

Prenatal Origin

There is well-established evidence for prenatal initiation of the leukemogenesis process in children (26-28). The early age of onset in childhood ALL (peak incidence in patients aged 2 to 5 years) is highly suggestive of a latency period that begins before the birth of the child. Moreover, molecular studies have shown that several of the common chromosomal translocations, mainly $t(4;11)MLL/AF4$ in infant ALL and $t(12;21)ETV6/AML1$ in childhood B-cell ALL, occur *in utero*. First, retrospective DNA screening of archived neonatal blood spots revealed the presence of the fusion genes in patients that developed ALL later on between the ages of 5 months to 2 years (26, 29). And concordance of ALL in identical, monozygotic twins – ranging from 50% in twin pairs diagnosed before the age of 1 to about 5% for older ages – has been shown to result from intraplacental metastasis of leukemic clones that initiate in one identical twin and are passed to the other *in utero* through their shared blood system (30). This was shown by mapping the unique genetic breakpoints in translocations which were shown to be shared among affected twin pairs (24). Moreover, screening of newborn cord blood revealed that 1% of babies carry the *ETV6/AML1* fusion (31), a frequency much higher than the actual prevalence of leukemia (about 1 in 10 000), providing a proof-of-principle that chromosomal translocations occur prenatally and are likely insufficient to cause leukemia. Other subtypes of ALL are expected to be initiated prenatally as well, but lack of genetic markers such as gene fusions precludes their identification.

Cooperating Oncogenic Lesions

While chromosomal alterations play an important role in driving the leukemic process by affecting molecular pathways that halt lymphoid progenitor cell differentiation and promote cell proliferation and survival, they are incapable, on their own, of causing full leukemic transformation (32). The observation that these abnormalities are often detected years before the onset of leukemia and

the use of experimental models to show that they do not alone result in leukemia (33) suggest the need for additional cooperating genetic mutations for full leukemic transformation. Greaves suggested that a minimum of two steps are required for leukemia development; the first genetic insult is expected to occur *in utero* during pregnancy, in perinatal or at the very least in early infancy, followed by a second postnatal genetic insult, that is required to induce overt leukemia (Figure 3).

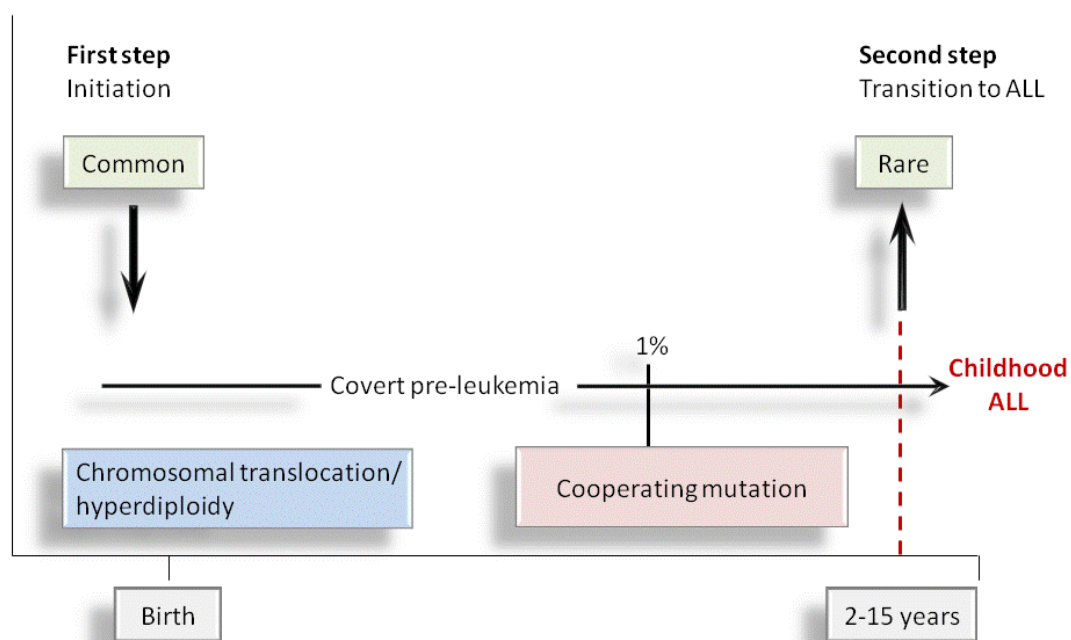


Figure 3. Greaves' multi-stage model for the development of childhood ALL

The most frequent genetic lesions in ALL are generated during normal fetal development and are fairly common events (31). Transition between covert pre-leukemia and covert clinical leukemia occurs in only 1% of the cases and involves at least a second hit involving cooperating mutations that occur postnatally.

Figure adapted from Greaves, 2006 (34).

Recent effort has gone into identifying the full complement of somatically acquired genetic alterations that contribute to acute leukemogenesis in children (35, 36). In addition to known recurrent chromosomal translocations, high frequency DNA copy number alterations, loss-of-heterozygosity events, deletions, amplifications and point mutations in genes involved in lymphoid differentiation, as well as in other critical cancer-related pathways such as cell cycle regulation, apoptosis, tumor suppression and xenobiotic metabolism have been identified in both B- and T-lineage ALL (37-39). Epigenetic analysis revealed that recurrent inactivation of tumor suppressor genes through DNA hypermethylation also contributes to oncogenesis in ALL (40, 41).

The recurrent chromosomal abnormalities observed in childhood ALL define unique subtypes of the disease and together with cooperating genetic and epigenetic alterations shed light on the pathogenesis of ALL in children (35, 42). Many of these alterations are associated with disease outcome and have helped guide risk-stratified treatment regimes contributing to the success of modern day therapy in curing childhood ALL. But while it is likely that these distinct prognostic subgroups also reflect divergent disease etiologies, very little is known still of the underlying causes of childhood ALL.

Risk Factors

Acquired genetic changes (somatic mutations) are involved in the molecular pathogenesis of childhood ALL and its progression but do not explain the origin of the disease. It is the more elusive inherited germline variants that influence disease susceptibility, likely through the modulation of the response to environmental exposures. But while it is clear that both environmental and genetic factors have roles in the development of ALL in children, studies have thus far been inconsistent or incapable of providing unequivocal evidence for

implication in the etiology of ALL and the causes of the disease remain largely unknown.

Evidence that childhood ALL has a genetic component stems in part from the fact that it is associated with other predisposing genetic syndromes. Inherited disorders such as Down's syndrome, Bloom's syndrome, ataxia-telangiectasia, and Nijmegen breakage syndrome are of the few established risk factors for childhood ALL however they account for a trivial proportion of cases (<5%) (43). Familial aggregation of childhood ALL is also rare and only a few pedigrees transmitting ALL have ever been recognized (43). There are five published reports of multigenerational ALL, possibly consistent with autosomal dominant inheritance, yet no clear susceptibility gene culprit has ever been identified within these families, and the small size and number of these pedigrees could be reminiscent of chance clustering rather than familial aggregation (43-46). Furthermore, very few reports of families with multiple affected children have been reported (47, 48), the incidence of childhood ALL among non-twin siblings of probands is at most only weakly increased and could even be decreased (49-50), and long-term follow up studies of childhood ALL survivors indicate no increased risk of malignancy in their offspring as compared to the general population (51), not what one would expect if familial and highly penetrant leukemia susceptibility genes were involved in disease etiology. Thus, as opposed to other forms of hereditary cancers, such as certain forms of breast, prostate and colorectal cancer for example, that are inherited in a Mendelian fashion and are associated with highly penetrant germline mutations, childhood ALL is sporadic and inherited genetic susceptibility is multifactorial and likely involves the co-inheritance of multiple low penetrance variants that do not give rise to clear-cut familial patterns of inheritance. And unlike the search for "the" gene that causes a Mendelian disorder, many different inherited susceptibility genes and many different environmental risk factors are likely going to be involved in the etiology of childhood ALL.

Environmental Exposures

In conjunction with Greaves' multi-stage model for childhood ALL development, there are three critical windows during which exposure to environmental risk agents could potentially influence leukemogenesis; these are before conception, *in utero* during pregnancy, and after birth (Figure 4). Given that ALL is not a hereditary cancer in terms of a simple Mendelian inheritance, the notion that the mother and/or father may play an important role in ALL development in the offspring is somewhat counterintuitive, unless we think of it in terms of exposure and increased mutation burden. Exposure to carcinogenic agents during gametogenesis can incur germline mutations in the gametes of the parents preconception that are passed down to the offspring and could lead to increased genetic instability postconception. The role of the father in preconceptional ALL risk may be more important than the mother since spermatogenesis occurs throughout the entire lifetime of the male, offering greater opportunities for mutations to occur, whereas females bear their oocytes from birth. During pregnancy however, the mother may play a crucial role in disease development as she provides the prenatal environment and can influence her offspring's risk of disease through environmental exposures passed to the fetus via the placenta or through the effects of her own genes that can directly influence the intrauterine milieu (52). Finally, postnatal exposures of the child both directly and indirectly through the mother, for example in breast feeding, could be important determinants of childhood ALL.

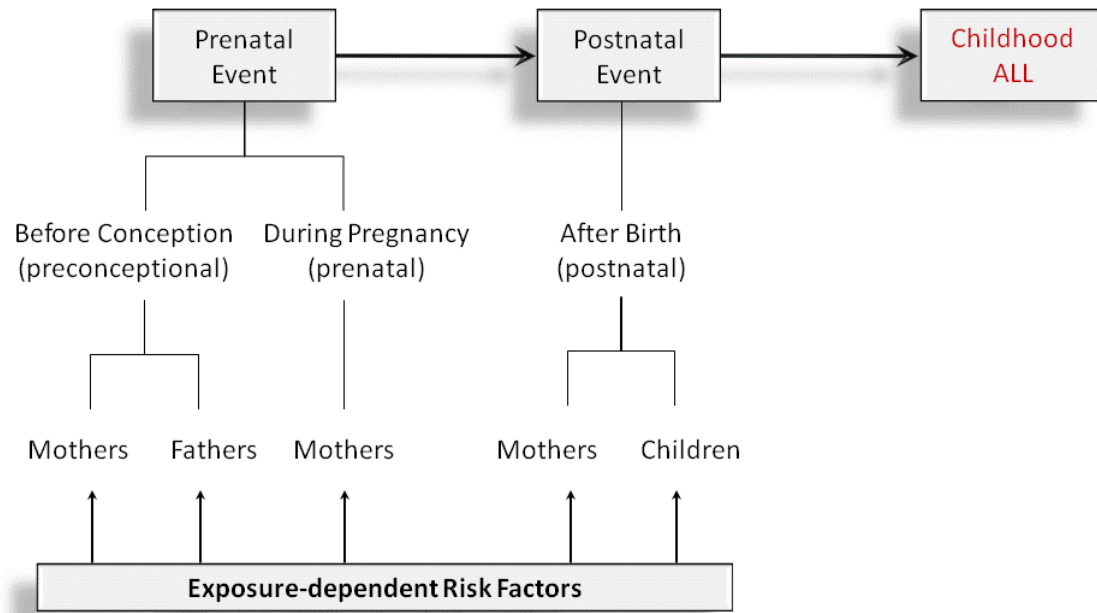


Figure 4. Critical windows of exposure for childhood ALL

Potential relationship between childhood ALL-inducing events and critical periods preconception, *in utero* and early after birth, during which exposure-dependent risk factors could influence disease susceptibility. In concordance with the multi-stage model for ALL development, both mothers and fathers could potentially be involved in the initiating genetic event while mothers could contribute to the child's postnatal exposure-dependent risk through breastfeeding. The model can also be extended to include genetic risk factors (at the level of the child, mother and/or father) given that response to the environment is modulated by genetic components such as those involved in xenobiotic metabolism, DNA repair and cell cycle regulation.

Figure adapted from Kim et al., 2006 (53).

Table 3 lists significant window-specific environmental exposures that have been associated with childhood ALL in the literature. The main classes of exposure-dependent risk factors identified for childhood ALL include ionizing radiation (preconception, *in utero* and postnatal), exposure to electromagnetic fields, chemicals (e.g. hydrocarbons such as benzene found in cigarette smoke, gasoline, solvents, paint thinners, air pollution and pesticide exposure either directly or via parental exposure), parental alcohol, cigarette and drug use, parental occupation, and certain dietary components. However contradictory results have been reported for most of these risk factors. Ionizing radiation appears to be the only significant environmental risk factor identified to date; most others (for example parental cigarette smoking and alcohol consumption) have been inconsistently associated with childhood ALL and their role in leukemogenesis remains very controversial. This stems mainly from the fact that exposures to environmental factors are often difficult to measure and assess and can suffer from high levels of uncertainty due to recall bias.

Table 3. Exposure-dependent risk factors reported to be significantly associated with childhood ALL

Exposure window	Exposure type	Risk ^a	Odds ratio (sample size ^b); reference
Preconception			
Fathers	Occupational exposure		
	Plastic materials (polystyrene)	+	1.4 (1842 vs. 1986); (54)
	Driving, exhaust fumes, inhaled particulate hydrocarbons	+	13-1.4 (1461 vs. 2922); (55)
	Electromagnetic fields	+	2 (56)
	Radiation exposure, X-rays	++	1.9-2.6 (184 vs. 368); (57) 2.2-3.8 (191 vs. 382); (58)
	Cigarette smoking	++	1.6 (203 vs. 406); (59) 3.8 (110 vs. 110); (60)
	Alcohol consumption	+	1.4 (491 vs. 491); (61)
	Medication/drug use		
	Amphetamines or diet pills	+	2.2 (1842 vs. 1986); (62)
	Mind-altering drugs (marijuana)	+	1.3 (1842 vs. 1986); (62)
Mothers	Occupational exposure		
	Solvents/hydrocarbons (carbon tetrachloride)	+	1.8 (1842 vs. 1986); (54)
	Paints or thinners	+	1.6 (1842 vs. 1986); (54)
	Household pesticides	+	1.7 (135 vs. 135); (63)
	Medication/drug use, oral contraceptives	+	1.3 (519 vs. 507); (64)
	Food consumption		
	Vegetables	-	0.5 (138 vs. 138); (65)
	Protein	-	0.4 (138 vs. 138); (65)
Prenatal			
Mothers	Occupational exposure		
	Solvents/hydrocarbons (freon, gasoline)	++	1.6 (1842 vs. 1968); (54) 1.7 (184 vs. 368); (57) 1.7-2.3 (790 vs. 790); (66)
	Paints or thinners	++	1.7 (1842 vs. 1986); (54) 3.2 (184 vs. 368); (57) 2.4 (519 vs. 507); (67)
	Pesticides	+	3.5 (184 vs. 368); (57)
	Organic dust (cotton, wool, synthetic fibers)	+	5.5 (128 vs. 128); (68)
	Electromagnetic fields	+	2.5 (491 vs. 491) (69)
	Household exposure		
	Pesticides	++	1.7-1.8 (491 vs. 491); (70) 2.3 (135 vs. 135); (63)
	Paints	+	1.7 (640 vs. 640); (71)
	Radiation exposure, X-rays	+	2.2 (519 vs. 507); (64)
	Electric blanket, heating pad, humidifier	+	1.4+1.6 (640 vs. 640); (72)

	Sewing machine	-	0.8 (640 vs. 640); (72)
	Cigarette smoking	--	0.7 (203 vs. 406); (59) 0.7; (73)
	Alcohol consumption	+	1.4 (203 vs. 406); (59)
		-	0.7 (491 vs. 491); (61)
	Medication/drug or supplement use		
	Vitamins	-	0.7 (1842 vs. 1986); (62)
	Iron or folate	--	0.4 (83 vs. 166); (74) 0.9 (1842 vs. 1986); (62)
	Antihistamines	+	1.3 (1842 vs. 1986); (62)
	Oral contraceptives	+	1.5 (1842 vs. 1986); (75)
	Pregnancy-maintaining drugs	+	1.9 (519 vs. 507); (64)
	Teratogenic medication (CNS depressants)	+	1.3-1.4 (789 vs. 789); (76)
	Antibiotics	+	1.5 (477 vs. 484); (77)
	Food consumption		
	Vegetables	-	0.8 (131 vs. 131); (78)
	Fruit	-	0.7 (131 vs. 131); (78)
	Fish and seafood	-	0.7 (131 vs. 131); (78)
	Sugars and syrups	+	1.3 (131 vs. 131); (78)
	Meat and meat products	+	1.3 (131 vs. 131); (78)
Postnatal			
Mothers	Occupational exposure, plastic (polyvinyl chloride)	+	2.2 (1842 vs. 1986); (54)
	Alcohol consumption	-	0.5 (491 vs. 491); (61)
Children	Environmental exposure		
	Neighboring repair garages/gas stations (benzene)	+	3.6 (240 vs. 280); (79)
	Chernobyl accident (radioactive contamination)	+	13.1 (98 vs. 151); (80)
	High voltage power lines (electromagnetic fields)	++	2.00; (81) 4 (101 vs. 412); (82) 1.69; (83)
	Household exposure		
	Pesticides	++	1.4-1.8 (491 vs. 491); (70) 1.7 (135 vs. 135); (63)
	Artwork (organic solvents)	+	4.1 (640 vs. 640); (71)
	Electrical appliance usage (electric blanket, hair dryer, video game machines)	+	1.6-2.8 (640 vs. 640); (72)
	Medication/drug use (chloramphenicol)	+	1.8-10.7 (184 vs. 368); (57)
	Supplementary oxygen exposure	+	1.9 (603 vs. 3015); (84)
	Radiation exposure, X-rays	++	1.6 (491 vs. 491); (85) 1.5 (701 vs. 701); (86)
	Trihalomethanes in drinking water	+	9.13 (491 vs. 491); (87)
	High birth weight (>3800 g)	++	1.3 (1905 vs. 9525); (88) 1.8 (1455 vs. 816); (89) 1.7 (603 vs. 3015); (84) 2.5 (181 vs. 362); (90) 2.2 (83 vs. 830); (91) 3.8-4.6; (92)

Infection ^c		
Early infection (in first 4 years of life)	--	0.4 (124 vs. 248); (93) 0.1-0.8 (408 vs. 567); (94)
Ear infection	-	0.3 (294 vs. 376); (95)
Roseola/fever and rash (first year of life)	-	0.3 (98 vs. 228); (96)
Allergy history	--	0.6-0.7; (97) 0.5-0.6 (1130 vs. 2957); (98) 0.6 (255 vs. 760); (99)
Vaccination ^c		
Haemophilus influenza type b (Hib)	--	0.6 (439 vs. 439); (100) 0.8 (282 vs. 409); (101)
Bacille calmette-guérin (BCG)	-	0.1 (63 vs. 126); (102)
Measles	-	0.2 (63 vs. 126); (102)
Household density (>1 person/room) ^c	-	0.6; (103)
Daycare attendance ^c	--	0.5 (490 vs. 491); (77) 0.7 (408 vs. 567); (94) 0.4 (294 vs. 376); (95)
Breastfeeding ^c	--	0.7 (491 vs. 491); (77) 0.8 (1744 vs. 1879); (104)

Only significant exposure-dependent risk factor associations are shown. This is not an exhaustive list; other postulated exposure risk factors for childhood ALL may have been investigated, inconclusive results are not shown.

^a Increased risk (odds ratio > 1) in one (+) or more (++) studies; decreased risk (odds ratio <1) in one (-) or more (--) studies.

^b The total number of childhood ALL cases versus healthy controls analyzed in each reference.

^c Thought to be infection-related risk factors.

Table adapted from Kim et al., 2006 (53).

Despite evidence that leukemogenesis is initiated *in utero*, high birth weight is one of the few birth-related factors that has been linked to ALL in children. Up to a 26% increase in ALL risk is associated with each kg increase in birth weight, which can perhaps simply be explained by the fact that larger babies have a higher number of lymphoid cells and therefore a higher number of cells at risk of leukemic transformation (88). On the other hand higher birth weight might indicate higher levels of circulating growth hormone which may induce proliferative stress on the bone marrow and indirectly be linked to leukemia (105). Other causative factors that have been proposed for childhood ALL include maternal (<20 years and >35 years) and paternal (>40 years) age, maternal reproductive history (miscarriage, abortion), high birth order (fourth born), long birth intervals (>5), race (white, Hispanic) and gender (male) however their associations with the disease remain highly speculative and the mechanisms through which they influence disease risk are unclear (53).

Moreover, it has been hypothesized that childhood leukemia could be caused by infection-related factors. An aberrant immunological response to infection at a vulnerable time in the child's life when lymphoid cell proliferation is high could render the child more susceptible to genetic insults and to leukemia. The possibility that ALL may have an infectious etiology is supported by the higher prevalence of leukemia in modern, wealthy societies and the appearance of clusters of childhood ALL cases in small residential communities (106-108). Two infection-based models for ALL development have been proposed: Kinlen's population mixing hypothesis and Greaves' delayed-infection hypothesis. Kinlen postulated that clusters of ALL could result from prolonged population isolation and subsequent exposure to a common but otherwise non-pathogenic infection due to population mixing (109). Whereas Greaves suggested that individuals carrying pre-leukemic clones (Figure 3) that have spent their early years of life coddled in a sterile environment may exhibit a pathological response to a subsequent delayed exposure to common infections (34). While some

epidemiologic data do support the immune-response hypothesis – daycare attendance, increased household density, higher number of recorded common infections in early life, as well as breastfeeding and vaccination have all been shown to reduce the risk of ALL suggesting that increased social contact and potential exposure to infection and immune stimulation in early life protect against ALL (Table 3)– no causal infectious agents have yet been identified.

Genetic Susceptibility Factors

How these exposures affect an individual and their susceptibility to disease relies largely on their genetic makeup. Human phenotypic variation, be it in risk to disease, response to the environment, or with regard to physical characteristics such as height, is influenced by inherited differences in DNA sequence. Interindividual variation in the susceptibility to childhood ALL bears no exception. Assuming that genes modulate variability in the responses to exogenous and/or endogenous factors, they could thereby also influence an individual's risk of cancer. Childhood ALL is a complex disease in which both genes and the environment interact to confer a variable degree of risk on people who inherit predisposing genetic variants. Genetic susceptibility refers to an inherited increase in the risk of developing a disease, passed down from parental generations in the form of germline genetic variation. While the parental generation may not be at increased risk of disease, the combination of polymorphic variants they bestow upon their offspring could render the child more or less susceptible to developing disease. In order to unravel the complex etiology of childhood ALL one must therefore start by identifying the inherited genetic changes that influence an individual's risk of disease.

To date, studies of genetic susceptibility to ALL have focused on the affected child and on common genetic variation in genes involved in biological pathways

such as folate metabolism, immune function, xenobiotic metabolism (including membrane transport, detoxification and biotransformation of drugs and chemicals), oxidative stress response, and DNA repair, under the presumed hypothesis that inherited genetic variants in genes functioning along these pathways could modify response to exposure-dependent risk factors and lead to genetic instability in lymphoid progenitor cells and thus influence a child's risk of developing ALL. Table 4 shows a summary of the genetic susceptibility factors shown to significantly modulate childhood ALL risk.

These hypothesis-driven approaches are based on our imperfect understanding of the biological processes involved in leukemogenesis and often yield associations that are difficult to replicate. And despite two recent large-scale association studies that have convincingly vindicated the role for inherited genetic variation in childhood ALL predisposition (110, 111), the genetic component of childhood ALL remains largely undefined.

Table 4. Genetic risk factors reported to be significantly associated with childhood ALL

Pathway, Gene	DNA variant	Risk	Odds ratio (sample size); reference
Folate metabolism			
<i>MTHFR</i>	677C>T	--	0.4 (71 vs. 71); (112) 0.4 (52 vs. 88); (113)
	1298A>C	-	0.4 (270 vs. 300); (114)
<i>MTRR</i>	66A>G	-	0.8 (460 vs. 552); (115)
Immune function			
<i>HLA</i>	DPB1*0101	-	0.7 (982 vs. 864); (116)
	DPB1*0201	+	1.8 (982 vs. 864); (116)
	DPB1*0402	+	1.3 (982 vs. 864); (116)
	DQA1*0101/*0104	+	2.3 (60 vs. 78); (117)
	DQA1*05	-	0.6 (60 vs. 78); (117)
Xenobiotic metabolism			
<i>GSTM1</i>	Null	++	1.8 (174 vs. 304); (118) 2.1 (118 vs. 188); (119) 2.2 (47 vs. 102); (120) 1.7 (107 vs. 320); (121)
<i>GSTP1</i>	*B 1578A>G	++	1.5 (278 vs. 303); (122)
<i>CYP1A1</i>	*2A 6235T>C	++	1.8 (170 vs. 299); (118) 2.6-6.2 (118 vs. 118); (119)
	m2 4889A>G	+	2.2-4.3 (118 vs. 118); (119)
<i>CYP2E1</i>	*5 1259G>C	+	2.8 (174 vs. 302); (123)
<i>MDR1</i>	3435C>T	++	2.5 (396 vs. 192); (124) 1.8 (113 vs. 175); (125)
	*4 No mutation	+	1.4 (155 vs. 306); (126)
<i>NAT1</i>	*4 No mutation	-	0.6 (176 vs. 291); (126)
<i>NAT2</i>	*5C 341T>C & 803A>G	+	3.1 (176 vs. 291); (126)
	*7B 282C>T & 857G>A	+	2.9 (176 vs. 291); (126)
Oxydative stress response			
<i>NQ01</i>	*2 609C>T	+	1.7 (174 vs. 323); (123)
	*3 465C>T	++	1.9 (174 vs. 323); (123) 6.4 (72 vs. 185); (127)
DNA repair			
<i>XRCC1</i>	Arg399Gln	+	1.9-2.4 (117 vs. 117); (128)

Only significant genetic risk factor associations as of December 2006, which coincides with the beginning of my study, are shown. This is not an exhaustive

list; other postulated genetic variants have been tested for association with ALL but are not shown here due to inconclusive results.

^a Increased risk (odds ratio > 1) in one (+) or more (++) studies; decreased risk (odds ratio <1) in one (-) or more (--) studies.

^b The total number of childhood ALL cases versus healthy controls analyzed in each reference.

Table adapted from Kim et al., 2006 (53).

Thus, while it is likely that both genetic and environmental factors are involved in the etiology of childhood ALL very few definitive risk factors have been identified. This is in part due to the complex nature of the disease and the fact that overall ALL risk is likely influenced by a complex interplay of multiple genetic, environmental, and perhaps infectious and other types of risk factors (e.g. birth weight, maternal age), acting at the level of the affected child as well as at the level of the parents. Identifying relevant interactions will likely be a challenging task however specific genetic variants shown to modify ALL risk associated with environmental exposures have already been reported (61, 70, 86, 87, 129); examples include *CYP2E1* and *GSTM1* with maternal alcohol consumption (61), and *CYP1A1* with pre- and postnatal exposure to pesticides (70). Furthermore, given the overwhelming evidence that childhood ALL is initiated *in utero*, attention must be paid to window-specific risk factors (Figure 4), particularly those affecting the prenatal period, in order to better understand the underlying mechanisms of the leukemic process; from a genetic susceptibility point of view this has seldom been done to date. The identification of risk factors for childhood ALL is critical to understanding disease etiology. And while we cannot change a person's genes, the identification of genetic risk factors could help identify high-risk populations and aid in the identification of modifiable environmental factors involved in leukemogenesis and provide new avenues for prevention.

DISSECTING THE GENETIC BASIS OF COMPLEX DISEASES

Complex diseases, such as cancers, have a tremendous public health impact and much effort has gone into understanding their causes in order to better treat them and to ultimately prevent their effects before they arise by identifying individuals who are at increased predisposition. Genetic epidemiology aims to detect patterns of disease in families and in human populations and identify the

inherited genetic variants, and jointly acting environmental factors, that are associated with disease susceptibility, to better understand the etiology of the disease and better control it (130). The completion of the first draft sequence of the human genome (131, 132) has brought the field of genetic epidemiology to a new era and has opened up many new resources for the study of the genetic determinants of disease; yet with new opportunities come new challenges. In this section I describe how these resources can be used to dissect the genetic basis of complex diseases.

The Landscape of Genetic Variation

Genetic variation comes in all shapes and sizes. Common polymorphisms (>1% in the population) include variable number tandem repeats (minisatellites, 0.1-20 kb; microsatellites, 2-100 nucleotides), large copy number variants, small segmental deletions/insertions (indels)/duplications, and single nucleotide polymorphisms (SNPs). SNPs are the most common form of variation, accounting for about 90% of all human genetic variation (Figure 5) (133). As of March 2010 a total of 23.4 million SNPs had been identified in the human genome (National Center for Biotechnology Information SNP database; dbSNP Build ID: 131, available from <http://www.ncbi.nlm.nih.gov/snp>) (134). SNPs occur on average once every 100-300 bps and are distributed throughout the genome with marked regional differences. SNPs can arise in coding regions of genes (exons), in the non-coding or regulatory regions of genes (introns, 5' and 3' UTR and promoter regions) or between genes in regions known as gene deserts. Many of these SNPs result in silent, neutral changes not affecting protein function, however some of these variants fall in coding regions or flanking regulatory regions and could have important consequences on molecular and physiological processes that may lead to disease (Figure 6).

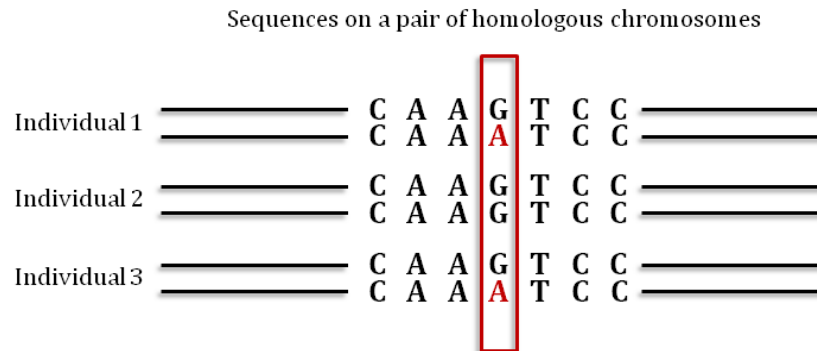


Figure 5. Single nucleotide polymorphism

SNPs are the most common form of genetic variation. They occur when a single DNA nucleotide (A, C, T or G) is substituted (e.g. here a replacement of a guanine by an adenine G>A). Almost all SNPs are biallelic, that is they have two sequence forms, or alleles, segregating in the population, however some have three (triallelic). In general the more common sequence form is called the major or wild-type allele, and the rare sequence variant is referred to as the minor allele.

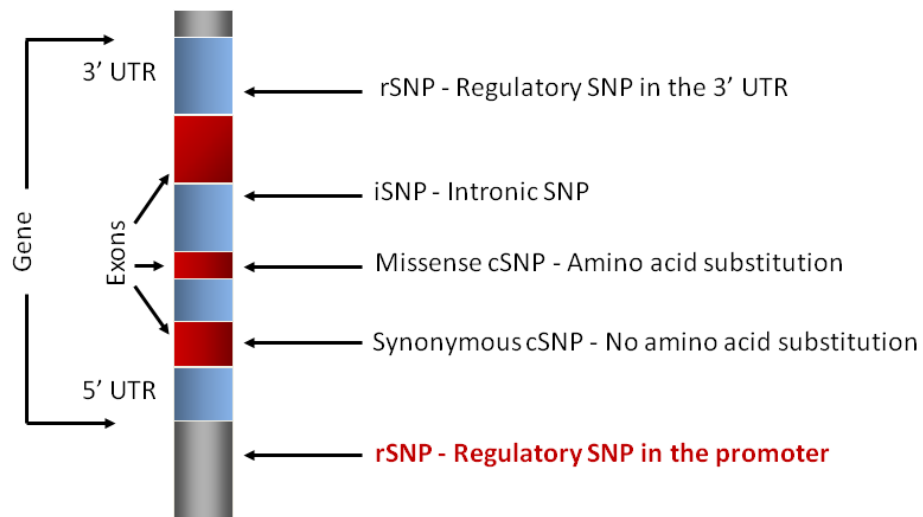


Figure 6. Gene-based single nucleotide polymorphisms

A SNP in a coding region of the gene (exons) may or may not affect the protein sequence encoded by the gene, due to degeneracy of the genetic code. A synonymous coding SNP (cSNP) will not change the amino acid sequence and is therefore silent however a nonsynonymous coding SNP can have a direct effect on protein structure and function by leading to a change in amino acid (missense cSNP) or by introducing a premature stop codon truncating the protein (nonsense cSNP). On the other hand, regulatory SNPs (rSNP) including intronic SNPs (iSNP), SNPs in the promoter and UTR regions can affect gene expression levels or alternative splicing.

While coding SNPs may have a direct effect on protein structure and function (qualitative change) by introducing changes in the amino acid sequence of the encoded protein or premature stop codons, it is thought that the softer, more subtle forms of regulatory variation (quantitative change) may play a more important role in complex diseases (135). Indeed, a large proportion of the susceptibility variants identified so far for a number of complex diseases including cancers, fall outside protein-coding regions. *Cis*-acting variants are commonly thought to involve regulatory elements such as promoters and enhancers which may lie immediately upstream of the gene, but can also be found hundreds of kilobases away (136). SNPs that lie within gene promoter regions can alter (change, destroy or create) recognition sites for transcription factor binding proteins (Figure 7). As a result, different transcription-regulatory proteins may be targeted to the promoter affecting temporal and/or spatial expression of the gene, or binding can be increased or decreased affecting gene dosage. Regulatory SNPs that alter gene expression are an important source of interindividual phenotypic variation and could play an important role in disease (Figure 8). Evidence of this has recently been demonstrated in humans where it was shown that common *cis* variants affecting gene regulation can explain over 50% of the observed population variation in gene expression (137).

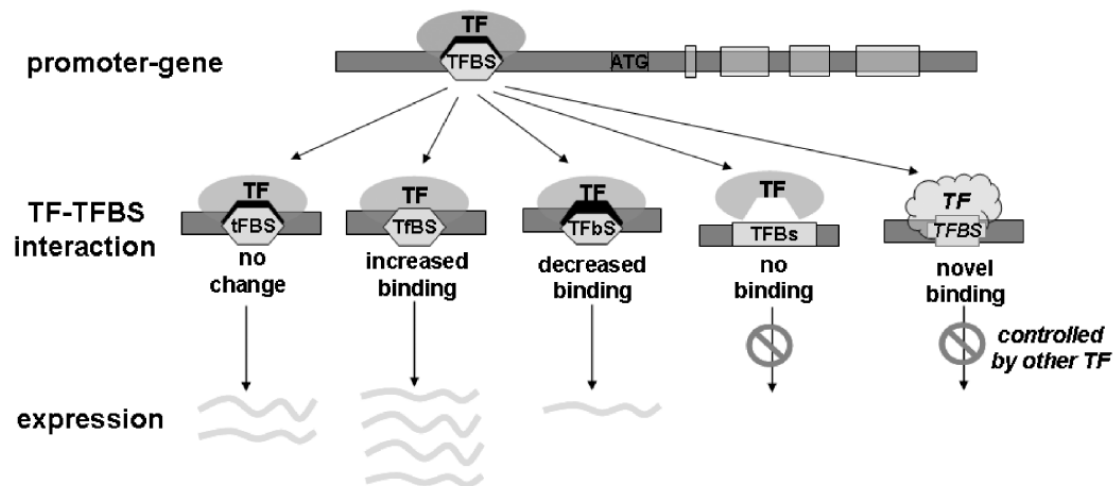


Figure 7. The impact of a SNP in a transcription factor binding site (TFBS)

A SNP that lies within a gene promoter can alter recognition sites for transcription factors (TF). In many instances, the SNP will not change TF binding affinity given that variation in the consensus sequence of the binding site is often allowed. However in some cases, a SNP may increase or decrease TF binding affinity leading to allele-specific gene expression. In rare cases, a SNP may destroy the TFBS or generate a novel binding site and consequently alter TF-induced control of gene expression.

Figure adapted from Chorley et al., 2009 (138).

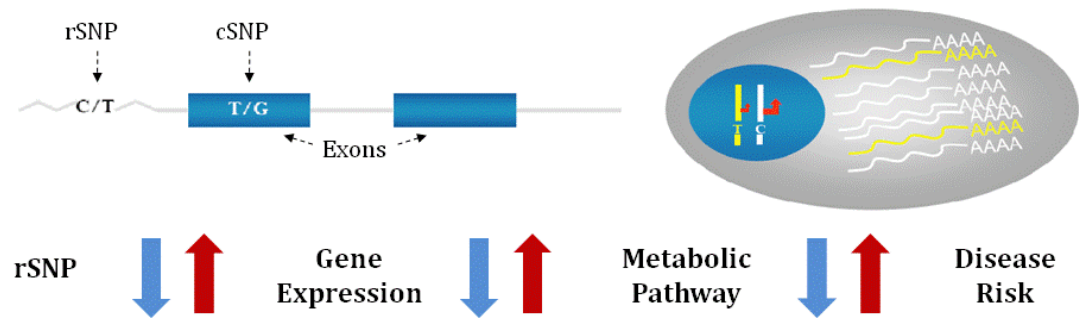


Figure 8. Rationale for rSNP discovery

Regulatory variation (rSNP) may have a stronger impact on phenotypic diversity as compared to coding variation (cSNP); this is supported by the increasing number of human disease associations that have been identified for which no deleterious coding variants can be found. Functional regulatory SNPs in gene promoter regions leading to allele-specific differences in gene expression could impact the overall outcome of the metabolic pathway in which the gene lies and thereby positively or negatively influence disease risk.

The allelic architecture of complex diseases (number, type and frequency of susceptibility variants) is expected to be as diverse as the diseases themselves. Due to their sheer abundance and the fact that they are easily amenable to experimental interrogation, common SNPs have been the markers of choice in genetic susceptibility studies. However many other forms of genetic variation including rare SNPs, DNA copy-number variants (CNV), large duplications, deletions, or inversions ranging from one kilobase to several megabases in size that can engulf one or many genes, are proving to be an important source of genetic variation and human phenotypic diversity (139, 140) and have been implicated in a number of diseases including predisposition to cancer (141). Equally important are inherited epigenetic modifications that can affect gene expression changes through DNA methylation and chromatin structure (142). Though it is likely that other forms of genetic variation contribute to interindividual variability in the susceptibility to childhood ALL, they are not the focus of this study and will not be discussed further here.

Identifying SNPs in the human genome has proven to be a relatively easy task compared to the complexity of determining which ones promote increased, or in certain instances decreased, susceptibility to disease. Considerable success has been achieved in finding genes responsible for Mendelian diseases (143). In contrast, we are still on the cusp of understanding genetic susceptibility to complex diseases. This lies mainly in the nature of the genetic variation involved in complex disorders. The human genome contains an estimated 20,000-25,000 genes (132) and in a hereditary Mendelian disorder only one of these is affected by a predisposing highly-penetrant mutation and is sufficient to cause disease. Highly penetrant mutations are for the most part rare but their effect on risk is important, and when inherited they virtually guarantee expression of the trait. However low-penetrance susceptibility variants are more common and are likely to be carried by a large proportion of the general population. Individually, they may only contribute modestly to complex disease risk but collectively, low-

penetrance susceptibility alleles could explain a large proportion of the disease. Hereditary cancers that exhibit familial clustering can involve both. For example in hereditary breast cancer, highly-penetrant mutations in the *BRCA1/2* gene explain a substantial proportion of the disease while a number of low-penetrant genetic variants interact with *BRCA1/2* to modify risk (144). It has long been thought that complex disease susceptibility is predominantly influenced by low-penetrance genes and that the allelic frequency spectrum for complex diseases tends toward multiple common variants (minor allele frequency, MAF $\geq 5\%$) with modest effects on disease risk (145, 146). This paradigm, known as the common-disease common-variant (CDCV) hypothesis (146-149) has long dictated how we search for susceptibility genes in complex diseases.

Genetic Association Studies

Traditionally two different approaches have been applied to identify disease susceptibility genes or loci: i) linkage analysis, which involves mapping susceptibility genes to locations along the chromosomes that are inherited with the disease within families; ii) association studies which entail investigating the relationship between a genetic polymorphism and disease in the population (150, 151) (Figure 9).

Linkage analysis relies on the principles of genetic recombination (crossing over of sister chromatids during meiosis) and identity by descent (IBD; refers to identical copies of the same ancestral allele) and always requires related individuals to identify genetic loci that cosegregate with the disease and that are physically linked to the putative susceptibility gene (i.e. on the same chromosome) (152). The rationale for linkage analysis is that related affected individuals will share more alleles at a particular locus if it is linked to the disease. Hence by finding genetic markers that are transmitted within families in

parallel with the disease, one can obtain an approximate idea of the chromosomal location of the true susceptibility gene. Different approaches include use of large family pedigrees or affected sib pairs to map tightly-linked markers to broad chromosomal regions which can then be narrowed down through additional fine-mapping in order to identify the disease gene more precisely.

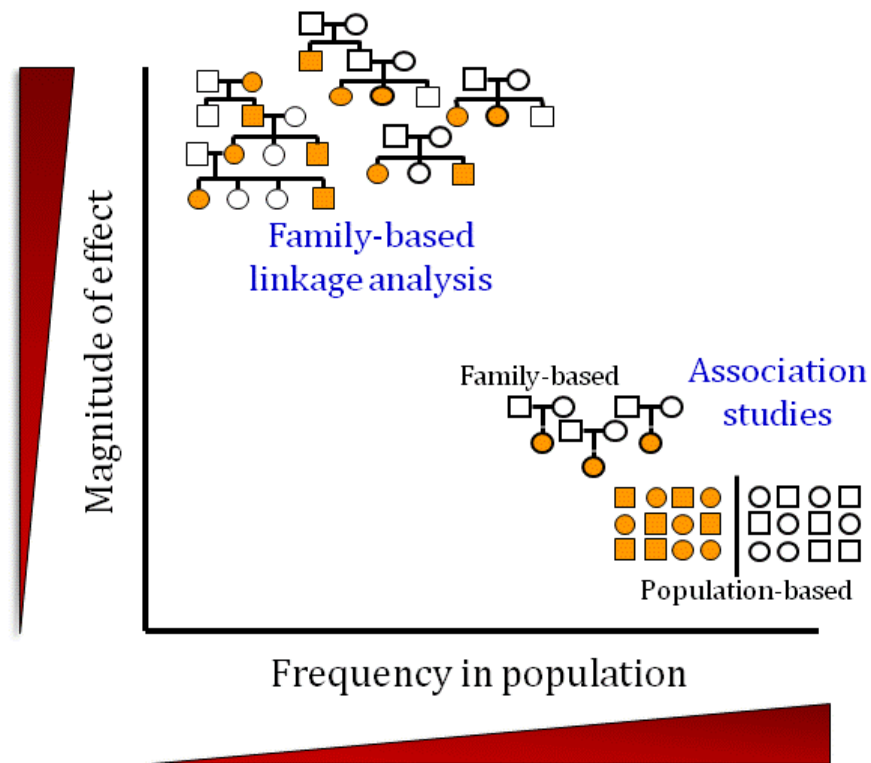


Figure 9. Disease mapping strategies

Linkage analysis requires related individuals and allows disease loci to be physically mapped along the genome. While linkage analysis is extremely useful for identifying high-impact disease predisposing genes, association studies are more advantageous for identifying common, low-impact susceptibility alleles. Association analysis can be performed either on unrelated individuals or using family-based designs.

While linkage studies have been extremely useful in identifying high-impact genes responsible for predisposition to hereditary cancers (153-156), they are less useful in identifying common variants with weak effects on disease susceptibility. In which case association studies are more commonly used to compare patterns of genetic variation between individuals with and without the disease or trait (Figure 10A). Most genetic association studies are conducted using either a population-based case-control or a family-based design.

Case-control Association Studies

In population-based association studies, unrelated individuals are used to identify alleles that are overrepresented in affected individuals (cases) relative to unaffected individuals (controls), which would indicate presence of a risk-enhancing gene, or that are underrepresented in affected individuals, which would indicate a protective effect against disease. In family-based association studies the control group is constructed from the genotypes of family members. Depending on the study design, different statistical methods are used to compare disease frequency (present or absent) between individuals with different risk factor characteristics (e.g. SNP alleles or genotypes) and to measure the strength of association.

Two types of measures for the effect of a genetic variant on disease are available: measures of risk and measures of odds of disease (150). However measures of risk between exposed and unexposed individuals (more commonly known as the relative risk, RR) can only be measured if the disease frequencies are incidence rates and these can only be obtained through cohort studies. In a cohort study a group of at-risk individuals (currently free of the disease under study) is identified, characterized in terms of their risk factors (e.g. SNP genotypes), and then followed over time to identify individuals that develop the

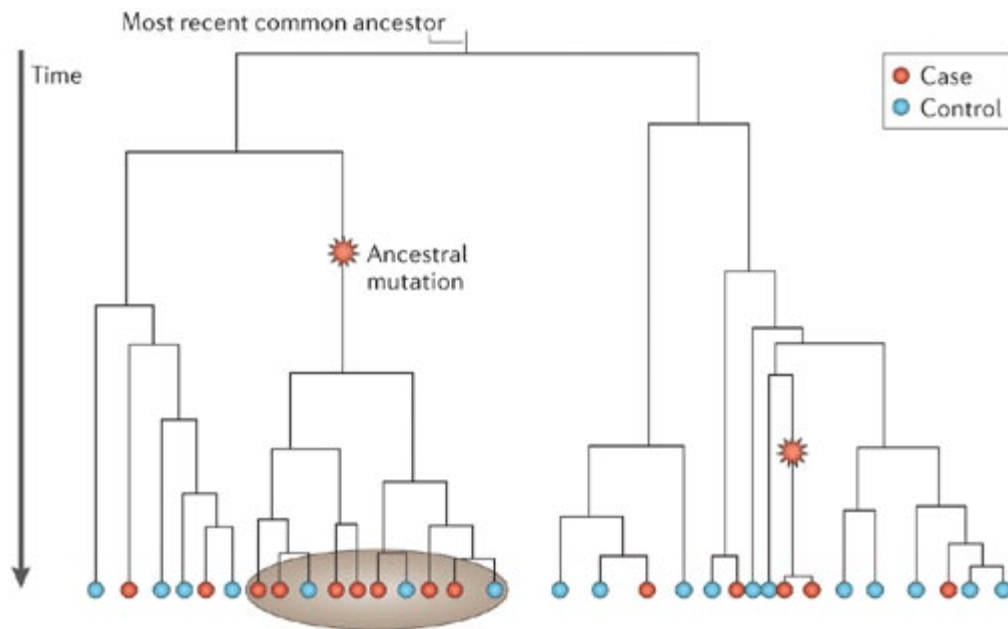
disease (152). Unless the disease of interest is common however, a cohort study could be a long and costly endeavor; which is why the preferred design for population-based association studies has become the case-control design. However given that study subjects are ascertained based on their disease status the measure of association in a case-control study cannot be based on disease frequency differences. Rather, the odds of carrying a given risk factor among affected cases is compared to the odds of carrying a given risk factor among unaffected controls (150). See Appendix I for details on measures of genetic association and their significance tests.

Linkage Disequilibrium

Evidence of “association” between disease outcome and the risk factor does not necessarily imply “linkage” and provides no information on the physical distance between the marker variant and the veritable disease locus. Rather, association testing often relies on the concept of linkage disequilibrium to identify variants that are correlated with the true susceptibility gene (Figure 10B). Linkage disequilibrium (LD) refers to the tendency of a pair of alleles to be associated with each other in the population more than would be expected by chance (157). Association studies rely on the fact that, although the causal variant may not be directly probed, its effects on disease may still be detectable through the polymorphisms with which it is in strong LD. However, whether the underlying rationale is that the selected polymorphisms directly affect the disease trait or that they are in LD with the true causal loci, it is often not possible to directly distinguish whether the gene variant under study is itself responsible for the observed effect on disease or if it is in LD with the true susceptibility gene; further functional validation is usually required to identify the true causal variant.

Association studies of common genetic variants are facilitated by the fact that the human genome is segmentally partitioned into a haplotype block-like structure composed of long segments of strong LD, interrupted by regions of low LD, known as hotspots for recombination (158). An individual who carries a particular SNP allele at one site often predictably carries a specific allele at another nearby variant site if the two SNPs are highly correlated (are in strong LD) and are in the same haplotypic block. And when a new mutation arises in the population, i.e. a SNP, indel or structural variant, it initially falls upon a specific haplotype associated with a particular combination of alleles. In association testing, haplotypes can be used to capture the correlation structure of the SNPs in the region in order to identify putative causal variants, and perhaps more importantly to capture the combined effects of tightly linked *cis*-acting variants (159).

A)



B)

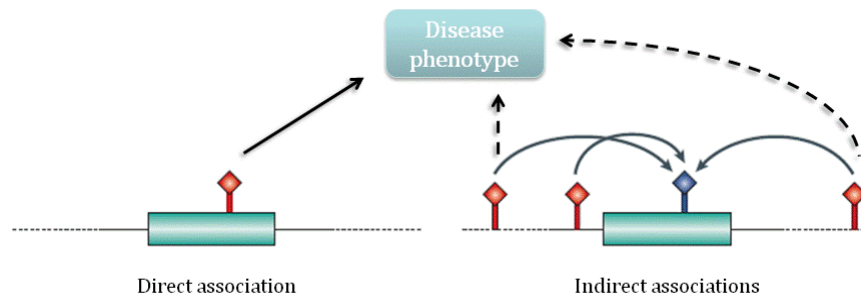


Figure 10. Rationale for association studies

A) Population-based association studies compare affected individuals (cases) to unaffected individuals (controls) that are presumed to be unrelated. Since we cannot trace transmission of the phenotype over generations (as in familial linkage analysis), we rely on correlations between current disease status or phenotype status with current marker alleles to infer statistical measures of disease association. Assuming that the causal disease locus has not been broken up by recombination over time, we should be able to identify groups of

individuals that share the susceptibility gene, provided they stem from a relatively recent common ancestor. See Appendix I for measures of association.

B) In an association study, the causal variant may not be directly interrogated. However if the unobserved causal locus is in strong linkage disequilibrium with one or more of the SNPs that were genotyped in the study, then its effects may still be detected. Identifying the true susceptibility variant often requires additional functional assays to provide evidence of disease causality. Typed markers are shown in red. In the left panel, the typed locus is also the true causal locus; in the right panel, the typed markers serve as proxies to the true causal locus (shown in blue).

Figure adapted from Balding, 2006 (160).

Family-Based Association Studies

An alluring feature of family-based association tests is the fact that they allow both linkage and association to be tested for. In other words, association will only be detected with markers that are both physically linked and in LD with the causal gene. A commonly used family-based design in genetic association studies is the family-based case-parent-trio design consisting of the affected proband and both parents. Though the trio design requires more resources, given that a total of three individuals need to be genotyped for each case, it is robust against biases such as population substructure that can lead to spurious associations in case-control studies (161). Cryptic population structure occurs when an apparently homogeneous population is in fact subdivided into subpopulations with varying allele frequencies such that cases may be disproportionately overrepresented by a subpopulation with higher frequency of the risk allele leading to a false-positive association. In family-based designs, the association test is performed on the family unit thus eliminating any problems related to population structure. Association is measured by comparing the observed number of alleles transmitted to the affected offspring with those expected under Mendelian transmissions (see Appendix I for details). An excess of risk alleles among affected offspring indicates that a disease susceptibility locus is in linkage and in linkage disequilibrium with the marker (162). Conversely, under-transmission of an allele is indicative of a protective effect against the disease.

Both population- and family-based designs have their uses in association testing. While case-parent-trios bar from unwanted population stratification, they can be more difficult to ascertain. It is usually much easier to recruit large numbers of unrelated cases and controls; though this may not be the case for pediatric disorders for which parents are usually available and more willing to participate. Family-based designs are of particular interest when studying

diseases with onset early in life, such as childhood ALL, in order to assess putative parental contributions to disease risk. However case-control designs are usually more efficient than family-based designs for detecting disease susceptibility genes (162-164). Main effects of environmental exposures on disease risk can only be estimated in a case-control design, while testing for interactions with environmental exposures could be facilitated in family designs (165, 166). Thus, both methods have their advantages and disadvantages and choosing an optimal study strategy relies largely on the specific questions being asked. It also often occurs that mixed sample types are ascertained, i.e. that parents of cases and unrelated control individuals are collected in parallel. In which case designs based solely on case-parent-triad data, or solely on case-control data, could lead to considerable losses in efficiency to detect associations if available genotype data are disregarded. In these instances additional approaches are required to combine the information provided from both study designs (e.g. (167-169)).

Hypothesis-Driven versus Hypothesis-Free Approaches

Association studies are typically of either two types: genome-wide or candidate gene. In a candidate gene approach, a specific subset of genes and variants (typically up to 50 SNPs) is selected to test for association with disease, based on biological plausibility or on functional evidence. Because candidate gene approaches are hypothesis-driven, they allow targeted evaluation of specific SNPs that can be chosen based on putative functional relevance (for example missense coding SNPs, or regulatory SNPs in promoter regions). Such targeted approaches could offer advantages in comparison to the genome-wide strategy, specifically for testing associations with low-frequency variants or in small study samples, and could facilitate identification of the causal variant given that SNPs are chosen based on functionality (170). However, the difficulty in candidate

gene approaches is of course identifying worthwhile candidates, given the low prior probability that any particular gene is involved in disease susceptibility.

At the other end of the spectrum is a more comprehensive genome-wide association approach which does not require any a priori information and allows surveying the entire genome for disease genes. With the advent of the International HapMap Project (171) and high-throughput genotyping technologies, it is now feasible and cost-efficient to capture most of the common genetic variation in a population using a limited, albeit large, number of SNPs. In a genome-wide association study (GWAS), up to a million SNPs can be assayed at a time, providing a powerful tool for investigating the genetic landscape of complex diseases. Due to their agnostic nature they are often referred to metaphorically as fishing expeditions. In other words, if we cast a large enough “net” over the human genome, we are bound to find something, regardless of whether or not it is truly significant. Thus, stringent cutoffs for declaring significance have been set for genome-wide studies and replication of an association in independent studies is a prerequisite to bar from the potentially high number of false-positives that could ensue from testing such a large number of markers. Despite the potential drawbacks, whole-genome association studies have proven to be invaluable for the study of the genetic basis of human disease and since their first appearance in 2005 (172), GWAS have identified susceptibility loci involved in a number of cancers including prostate cancer (173), breast cancer (174), lung cancer (175-177), colorectal cancer (178, 179), and recently childhood ALL (110, 111). However given that GWAS identify regions of association, extensive fine-mapping is required to refine the risk loci and identify functional risk alleles along the associated haplotypes (180). Another important limitation of GWA screening is its insensitivity to rare variation. This is mainly due to the fact that whole genome association scans rely on SNP and LD maps which are strongly biased towards common variation (mostly $MAF > 5\%$) under the prevailing paradigm that

complex diseases are caused by common genetic variants. Consequently, although GWAS are an important advance in the study of the genetic architecture of complex diseases, they are not without their technical challenges.

Finding the Missing Heritability

A decade after decoding the human genome (132, 181), surprisingly little is known of the genetic architecture of complex diseases. While it was thought that the human genome would be opened up like a book and the genetic components of common traits and diseases such as cancer would be readily divulged, this has not been the case. And despite much of the enthusiasm surrounding GWAS, they have yet to explain but very little of the inherited susceptibility for most of the diseases investigated to date (182). Complex diseases have proven to be utterly more intricate and multifaceted than anticipated. Many explanations have been offered for this missing heritability in complex traits including a larger number of variants with small effects that have yet to be identified (the so-called low hanging fruit as opposed to the higher impact loci that have so far been singled out by GWAS); rarer variants ($MAF < 5\%$) that aren't interrogated through many current candidate and GWA studies; structural variation such as copy number variants that have been paid little attention; more complex scenarios involving gene-gene and gene-environment interactions; non-coding epigenetic changes; phenotype ascertainment issues (182, 183). In the case of pediatric disorders, another level of complexity which may hinder identification of susceptibility genes is the possibility that the parents' genes, particularly maternal genetic effects, may contribute to disease risk in the offspring. Therefore in a study of the genetic determinants of childhood ALL, it is unlikely that interrogating single, inherited DNA polymorphisms is going to reveal all or much of the underlying complexities of the disease.

Finding the so-called “dark matter” of cancer susceptibility is a challenging task and while many other sources of missing heritability, such as rare variants and gene-environment interactions, are likely to explain part of the interindividual variability in the susceptibility to childhood ALL, they are beyond the scope of my research. In my thesis I focused on the possible contribution of gene-gene interactions, maternally-mediated genetic effects, and pathway-specific effects which may explain a proportion of the inherited risk of ALL among children.

Epistasis

Part of the complexity underlying the etiology of childhood ALL might be attributed to epistasis or gene-gene interaction which occurs when the effect of one gene is modified by the effect of another. The term interaction can be interpreted in several ways (184) but essentially, it refers to the departure from independence of the effects of different loci in the way that they combine to cause disease (185). In this way, statistical interaction can be defined as the departure of the observed risk of disease from some model (e.g. additive or multiplicative, depending on the scale on which the effects are measured) for the main effects of two or more independent factors (152). It has been argued that gene-gene interactions are a ubiquitous component of the genetic architecture of human diseases and that epistasis may even be more important than the main effects of any single, inherited susceptibility variant (186). A number of interacting loci have been shown to influence cancer risk including prostate cancer (187), bladder cancer (188, 189), adult myeloid leukemia (189) as well as childhood ALL (126). The difficulty however lies in our capacity to link statistical interaction to biological interaction, for one does not directly imply the other. Biological interaction is the result of physical interactions between proteins within a gene regulatory network or a biochemical pathway that alter a phenotype in an individual (184, 190). In contrast, statistical epistasis describes the relationship between multilocus genotypes and their combined effects in phenotypic variation in a population. Both perspectives of epistasis, though

derived from and leading to different research strategies, are thought to play an important role in the etiology of common human diseases but the extent to which statistical interaction implies underlying biological interaction is still very much disputed and experimental support is often required to relate the two (191).

Maternal Genetic Effects

The genetic component of early onset diseases has an added level of complexity involving not only the affected individual's inherited genotype but also parentally mediated genetic mechanisms (169). In particular, the mother plays a crucial role as not only genetic contributor but also as fetal environment. A maternal allele can damage the fetus through its effect on the intrauterine milieu, regardless of whether it is transmitted (52). For example maternal exposures to environmental carcinogens and the efficiency of the mother's detoxification system, determined by the mother's genotype, could negatively affect the fetal environment and lead to increased susceptibility to disease for the growing fetus. Moreover there is increasing evidence that the nutritional, hormonal and metabolic environment of the mother may permanently programme the structure and physiology of her offspring and that adverse influences *in utero* might increase the risk of disease later on in adult life (192, 193). The genetic risk associated with diseases originating *in utero* can therefore be influenced by the inherited genotype of the child and/or the prenatal environment, as determined by the maternal environment and her own genotype, or by joint maternal-gene and offspring-gene effects (194). Investigation into maternally-mediated genotype effects in the susceptibility to childhood ALL is therefore a prerequisite for unraveling the genetic complexities of this disease.

Candidate Cancer Pathways

There is increasing interest in the use of pathways or networks to help guide association testing (195, 196). Given that genes function within complex molecular networks, rather than look at a single gene at a time (or on the contrary at all the genes at the same time, such as in a GWAS), it could be more biologically relevant to interrogate genetic variants along a given pathway, or in interconnecting or even conflicting pathways, in order to gain further insight on the molecular mechanisms involved in disease etiology. Pathway-driven approaches to studying complex diseases have primarily been used to select promising candidates for association testing but seldom has the information been carried through to the analysis stage (e.g.(197, 198)). Methods for incorporating pathway information to guide statistical inference have been proposed (199-202) and can help not only in the identification of individual SNPs and entire pathways involved in disease, but also in the characterization of the underlying complexities of the disease process.

Cancer, such as childhood ALL, is the end result of a network of multiple interacting genetic and environmental factors affecting common molecular routes that all converge toward oncogenesis. Biological pathways known to contribute to cancer susceptibility include xenobiotic metabolism, oxidative stress response, DNA damage control, cell cycle regulation and apoptosis (Figure 11). Dysregulation of any one of these pathways can lead to increased mutational burden, increased genomic instability and ultimately promote oncogenesis. My thesis work focused on the cell cycle control and DNA repair pathways. Genes encoding key regulators of DNA damage response and cell cycle progression are known as the caretakers and gatekeepers of the genome and inherited polymorphisms in either type of gene can predispose an individual to cancer.

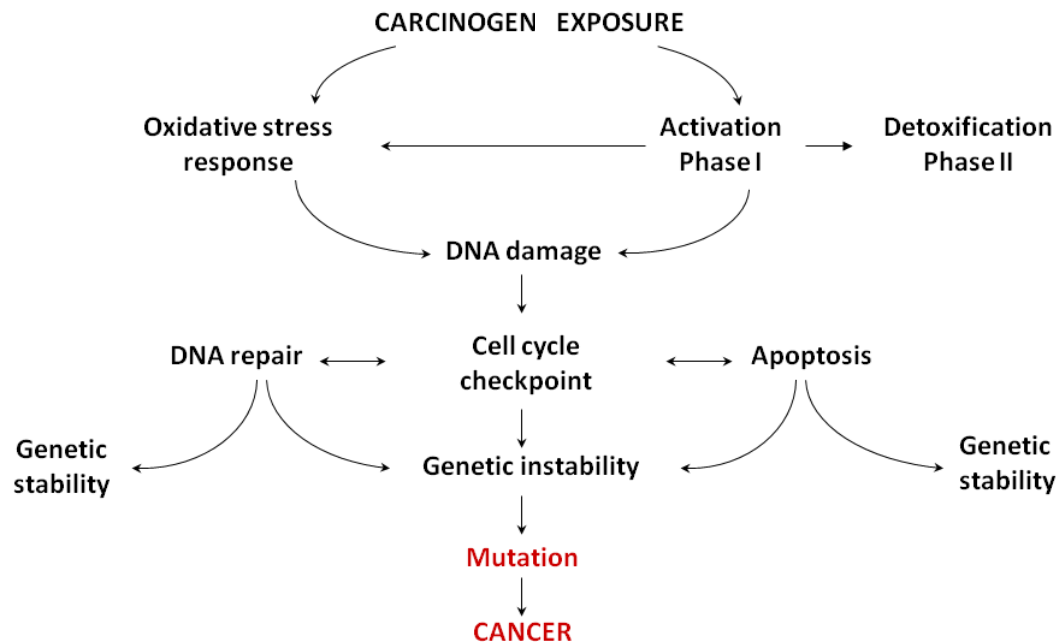


Figure 11. Cellular responses to environmental exposures and cancer susceptibility

Biological processes known to contribute to cancer susceptibility include those involved in modulating response to adverse environmental exposures : xenobiotic metabolism, oxidative stress response, DNA repair, cell cycle control and apoptosis. Dysregulation of these interconnected pathways can lead to increased genetic instability, increased mutation burden and eventually cancer.

The Cell Cycle and Cancer

At the heart of these interconnected pathways is the cell cycle regulatory network (203, 204). The proliferative advantage of a cancer cell arises from its ability to bypass critical cell cycle checkpoints in response to external signals; loss of cell cycle control is a hallmark of oncogenesis. The cell cycle is a highly-ordered and tightly-regulated process that is divided into four sequential phases: G1 is the growth phase during which cells integrate external information derived from mutagenic stimuli and nutrient availability to prepare for DNA replication which occurs during S phase; during the G2 gap phase the cell prepares for the M (mitosis) phase during which chromosomes are partitioned equally into two identical nuclei and cytokinesis occurs producing two identical daughter cells. Regulated progression through the cycle is ensured by cell cycle checkpoints, the main function of which is to ensure that the integrity of the genome remains intact. The checkpoints involve complex networks of signaling proteins that can sense aberrant or incomplete cell cycle events, and through a series of signal transduction pathways, can engage effector proteins that either invoke cell cycle arrest allowing time for the problem to be resolved, or if the damage is irreparable, invoke programmed cell death. Concerted action of a vast number of signaling molecules is required to ensure strict regulation of the checkpoint function. The G1/S transition checkpoint is a critical point in the cell cycle during which the cell irreversibly commits to a new round of division, it is also a major sensor of DNA damage and its proper function is crucial in order to maintain genomic integrity (Figure 12). Not surprisingly, somatic mutations in key regulators of G1/S phase progression and changes in expression of their protein products have been reported in almost all human cancers, and germline polymorphisms have been associated with increased risk to a number of cancers, highlighting the importance of dysregulated cell cycle control in the initiation and progression of oncogenesis (204, 205).

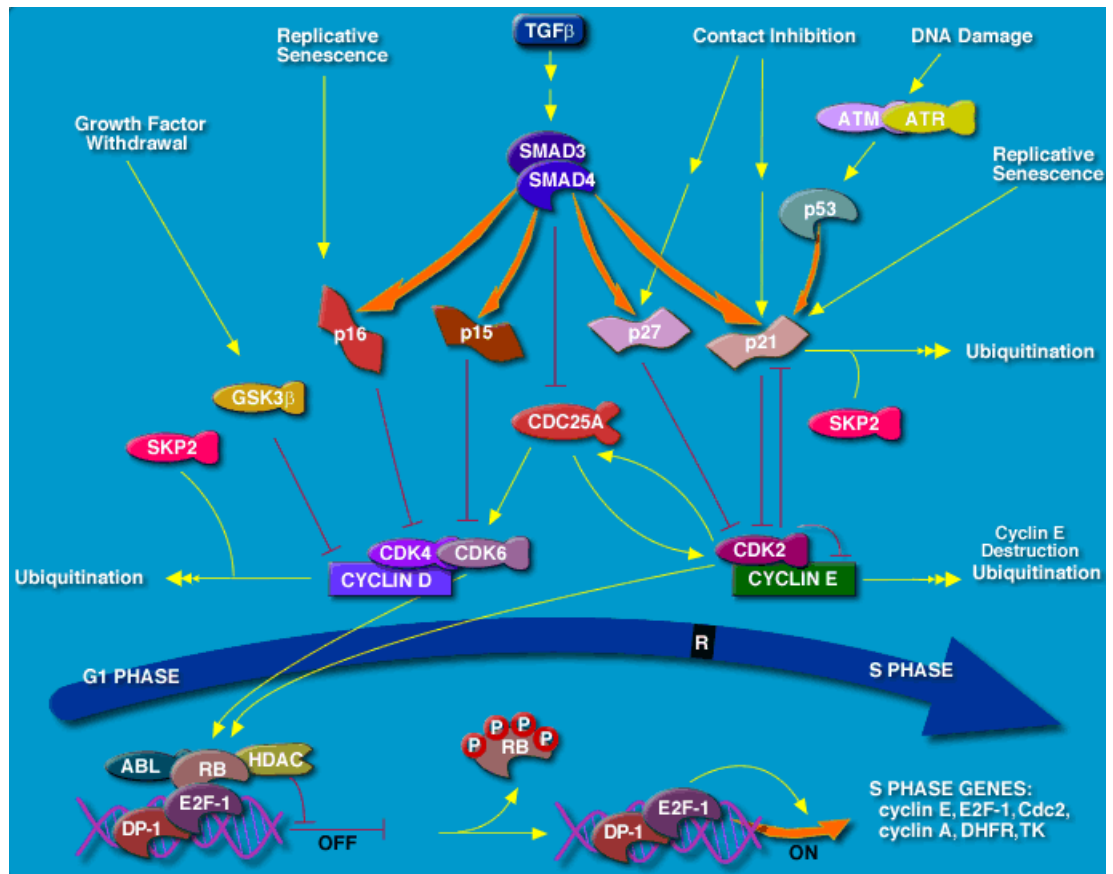


Figure 12. Illustration of the G1/S cell cycle checkpoint

The G1/S cell cycle checkpoint controls the passage of eukaryotic cells from gap phase G1 into the DNA synthesis phase (S). Two cyclin/cyclin-dependent kinase (CDK) complexes, CDK4/6-cyclin D and CDK2-cyclin E, and the transcription complex that includes RB and E2F are pivotal in controlling this checkpoint. During G1 phase, the RB-HDAC repressor complex binds to the E2F-DP1 transcription factors, inhibiting downstream transcription. Phosphorylation of RB by CDK4/6 and CDK2 dissociates the RB-repressor complex, permitting transcription of S-phase genes encoding for proteins that are required for DNA replication. Many different stimuli exert checkpoint control including the growth-inhibitory cytokine TGFβ, DNA damage, contact inhibition, and growth factor withdrawal. The first three act by inducing cyclin-dependent kinase inhibitors (CDKIs) of the INK4 (p16/CDKN2A and p15/CDKN2B) and Cip/Kip (p21/CDKN1A and p27/CDKN1B) families. TGFβ additionally inhibits the transcription of CDC25A, a phosphatase that activates the cell cycle

kinases. Cell cycle arrest in response to DNA damage is mediated through p53. Depending on the extent of the damage, p53 can trigger repair pathways or induce apoptotic cell death. The restriction point, R, is the crucial point at which the cell is irreversibly committed to undergo another round of division.

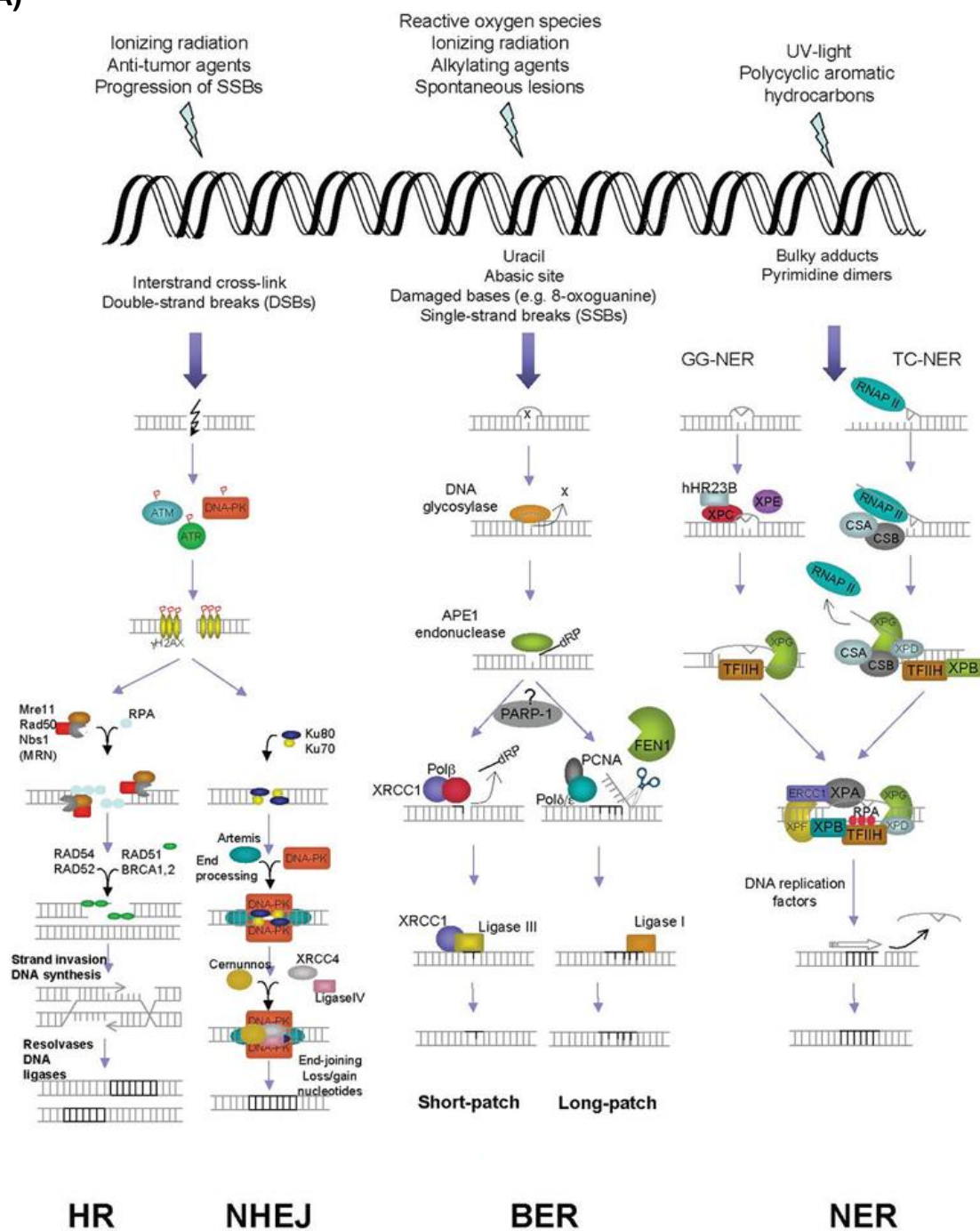
Figure adapted from BioCarta

(http://www.biocarta.com/pathfiles/h_g1Pathway.asp).

DNA Repair Pathways and Cancer

Cellular responses to DNA damage constitute one of the most important processes in cancer research, first because DNA damage induces cancer, second because it is used to treat cancer and third because DNA damage is responsible for most of the side-effects attributable to cancer therapies, including second neoplasias, hair loss, immune deficiency (206). The integrity of the genome is constantly being threatened by endogenous (e.g. reactive oxygen species caused by normal metabolism, DNA replication errors) and exogenous (e.g. UV light, ionizing radiation, and genotoxic agents such as benzene in cigarette smoke) stressors. In response to the various types of DNA lesions that can occur, a variety of different DNA repair systems are in place to protect the genome: excision repair, mismatch repair (MMR), and double-strand break repair (DSBR) (Figure 13). Impaired DNA damage responses can lead to genetic instabilities which result in increased mutation in the genome and ultimately give rise to oncogenic transformation. Genomic instability due to impaired DNA damage responses is a hallmark of all cancers. The DSBR pathway is particularly relevant to childhood ALL given that double-strand breaks are at the origin of the chromosomal translocations observed in a large proportion of leukemia patients (24). Moreover, the identification of a number of cancer predisposing syndromes linked to defects in DNA repair pathways support the hypothesis that DNA repair-related genes play a key role in cancer development and progression (207, 208). Interindividual variability in the response to carcinogens associated with variations in DNA repair capacity/efficiency have also been linked to polymorphisms in a number of DNA repair genes and modulated risk of developing a number of sporadic cancers (209, 210).

A)



B)

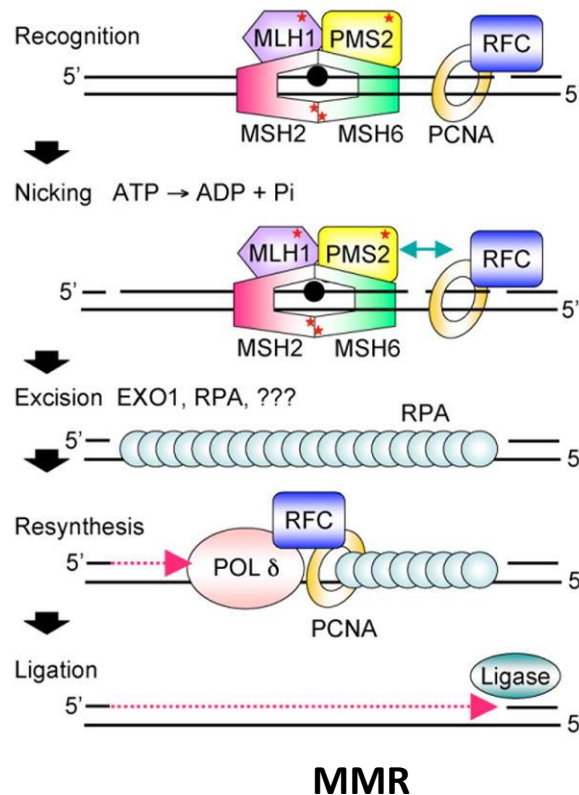


Figure 13. DNA repair mechanisms

A) Double-strand breaks (DSBs) can result from either exogenous factors such as ionizing radiation or genotoxic agents or from endogenous sources such as reactive oxygen species or spontaneously during replication or normal development of the immune system. Two major pathways exist for the repair of DSBs: homologous recombination (HR) and non-homologous end-joining (NHEJ) (211). HR uses homologous undamaged DNA from the intact sister chromatid and is therefore restricted to the S and G2 phases of the cell cycle. NHEJ functions in all phases by ligating broken, often noncompatible DNA ends, and is the dominant repair mechanism in mammalian cells. Deamination, oxidation, or alkylation of bases caused by oxidative stress, ionizing radiation or alkylating agents are repaired through base excision (BER). Pathway preference (long- or short-patch) depends on the type of lesion, the state of the cell cycle and the differentiation state of the cell (212). UV exposure induces bulky, helix-distorting lesions in the DNA that are repaired through the

nucleotide excision repair (NER) pathway (213). The two pathways of NER, global genome repair (GG-NER) and transcription-coupled repair (TC-NER), differ only in lesion recognition.

Figure adapted from Mostoslavsky, 2008 (214).

B) Mismatching of DNA bases can occur as a result of DNA polymerase errors or strand slippage during replication through microsatellite regions for example, and are usually repaired through the mismatch repair mechanism (MMR) (215).

Figure adapted from Hsieh and Yamane, 2009 (215).

OBJECTIVES OF THE PROJECT

The etiology of childhood ALL is likely explained by a combination of genetic susceptibility and environmental exposure during early development in fetal life and infancy. Assuming that genes modulate individual response to exogenous and/or endogenous factors, they would thereby also influence a child's risk of cancer. The goal of this study was to investigate the genetic basis of childhood ALL and to identify potential genetic risk factors. We postulated that a child's susceptibility to ALL is modulated, at least in part, by functional DNA variants in genes involved in the G1/S cell cycle checkpoint and in the cellular response to DNA damage. These carefully orchestrated processes are under strict regulatory control; it seems plausible that variation in gene dosage and activity of critical cell cycle control and/or DNA repair genes due to functional regulatory polymorphisms could have an important impact on the overall maintenance of genomic integrity and cell homeostasis, and thus influence disease.

For my doctoral research, I wished to further elucidate the mechanisms through which genetic variants in these biological pathways might affect disease susceptibility and attempt to explain some of the observed interindividual variability in the risk to developing ALL among children. Using a unique study design consisting of case-parent trios as well as unrelated pre-B ALL cases and healthy controls from the established Quebec childhood ALL cohort, I conducted a candidate gene association study to investigate the putative role of 12 cell cycle control genes (*CCND1*, *CDC25A*, *CDKN1A*, *CDKN1B*, *CDKN2A*, *CDKN2B*, *E2F1*, *HDAC1*, *MDM2*, *SMAD3*, *RB1*, *TGFB1*) and 7 DNA repair genes (*ATM*, *BRCA1*, *BRCA2*, *RAD51*, *XRCC4*, *XRCC5*, *XRCC6*) in childhood ALL susceptibility. The specific aims of this study were:

- 1) investigate individual main effects including single-site allelic/genotypic as well as multilocus haplotypic effects, of candidate genes on the susceptibility to childhood ALL;
- 2) investigate the combined effects of multiple genes/variants on disease risk through pathway-specific and gene-gene interaction effects;
- 3) assess the role of maternally-mediated genetic effects in the susceptibility to childhood ALL.

EXPECTED IMPACT OF THE PROJECT

In this project I attempted to dissect the genetic factors that shape interindividual variability in the susceptibility to childhood ALL. In doing so, I hope to be able to offer new insight into the process of leukemogenesis and possibly provide novel targets for risk-based disease management or even novel therapeutic targets for disease treatment. The significance of this project lies in the unique opportunity to better understand the impact of regulatory genetic variation in cell cycle and DNA repair genes on the risk of pre-B ALL. This is one of only a few studies to assess the impact of entire biological pathways on childhood leukemogenesis and is the only one to have addressed the role the mother's genes play in the development of leukemia among children. The identification of genes that modify a child's risk of ALL, together with a better understanding of how genes interacting in biological pathways are involved in more complex disease mechanisms, will allow further elucidation of the genetic architecture of this disease and perhaps help refine our understanding of pediatric cancers in general. Ultimately the goal of a genetic epidemiology study such as this is to contribute to the overall improvement of population health, and be able to impact on policy and planning in the health sector, particularly, with regards to this study, in the fields of pediatric oncology and preventive medicine.

REFERENCES

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, et al. Cancer statistics, 2008. *CA Cancer J Clin*. 2008 Mar-Apr;58(2):71-96.
2. Ries LAG, Percy CL, Bunin GR. *Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995*. Bethesda, MD: National Cancer Institute; 1999 (NIH Publication No. 99-4649).
3. Carozza SE, Chen VW, Wu X-C, Lehnher M, Derrick L, Riley R, et al. *Childhood Cancer Incidence in North America, 1988-1992*. In Howe HL (Ed.). *Cancer in North America, 1989-1993*. Sacramento, CA: North American Association of Central Cancer Registries, April 1997 Vol 1; V-1-13.
4. Steliarova-Foucher E, Stiller C, Lacour B, Kaatsch P. International Classification of Childhood Cancer, third edition. *Cancer*. 2005 Apr 1;103(7):1457-67.
5. Smith MA, Ries LAG, Gurney JG, Ross JA. *Cancer Incidence and Survival among Children and Adolescents: United States SEER Program 1975-1995*. Bethesda, MD: National Cancer Institute; 1999 (NIH Publication No. 99-4649).
6. Colby-Graham MF, Chordas C. The childhood leukemias. *J Pediatr Nurs*. 2003 Apr;18(2):87-95.
7. Passegue E, Jamieson CH, Ailles LE, Weissman IL. Normal and leukemic hematopoiesis: are leukemias a stem cell disorder or a reacquisition of stem cell characteristics? *Proc Natl Acad Sci U S A*. 2003 Sep 30;100 Suppl 1:11842-9.
8. Larsson J, Karlsson S. The role of Smad signaling in hematopoiesis. *Oncogene*. 2005 Aug 29;24(37):5676-92.
9. Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton DA, Gralnick HR, et al. Proposed revised criteria for the classification of acute myeloid leukemia. A report of the French-American-British Cooperative Group. *Ann Intern Med*. 1985 Oct;103(4):620-5.
10. Rothe G, Schmitz G. Consensus protocol for the flow cytometric immunophenotyping of hematopoietic malignancies. Working Group on Flow Cytometry and Image Analysis. *Leukemia*. 1996 May;10(5):877-95.
11. M. C. Kinney, Lukens JN. *Section 2. The Acute Leukemias, Classification and Differentiation of the Acute Leukemias* Wintrobe's Clinical Hematology 10th Revised edition: Lippincott Williams and Wilkins; 1998. p. 2209-40.

12. Pieters R. Infant acute lymphoblastic leukemia: Lessons learned and future directions. *Curr Hematol Malig Rep*. 2009 Jul;4(3):167-74.
13. Uckun FM, Sensel MG, Sun L, Steinherz PG, Trigg ME, Heerema NA, et al. Biology and treatment of childhood T-lineage acute lymphoblastic leukemia. *Blood*. 1998 Feb 1;91(3):735-46.
14. McLaughlin J, Dryer D, Logan H, Mao Y, Marrett L, Morrison H, et al., editors. *Canadian Cancer Statistics 2006*. Toronto: National Cancer Institute of Canada, 2006.
15. Tower RL, Spector LG. The epidemiology of childhood leukemia with a focus on birth weight and diet. *Crit Rev Clin Lab Sci*. 2007;44(3):203-42.
16. Parkin DM, Stiller CA, Draper GJ, Bieber CA. The international incidence of childhood cancer. *Int J Cancer*. 1988 Oct 15;42(4):511-20.
17. McNeil DE, Cote TR, Clegg L, Mauer A. SEER update of incidence and trends in pediatric malignancies: acute lymphoblastic leukemia. *Med Pediatr Oncol*. 2002 Dec;39(6):554-7; discussion 2-3.
18. Brenner H, Kaatsch P, Burkhardt-Hammer T, Harms DO, Schrappe M, Michaelis J. Long-term survival of children with leukemia achieved by the end of the second millennium. *Cancer*. 2001 Oct 1;92(7):1977-83.
19. Hudson MM, Jones D, Boyett J, Sharp GB, Pui CH. Late mortality of long-term survivors of childhood cancer. *J Clin Oncol*. 1997 Jun;15(6):2205-13.
20. Mertens AC, Yasui Y, Neglia JP, Potter JD, Nesbit ME, Jr., Ruccione K, et al. Late mortality experience in five-year survivors of childhood and adolescent cancer: the Childhood Cancer Survivor Study. *J Clin Oncol*. 2001 Jul 1;19(13):3163-72.
21. Forestier E, Schmiegelow K. The incidence peaks of the childhood acute leukemias reflect specific cytogenetic aberrations. *J Pediatr Hematol Oncol*. 2006 Aug;28(8):486-95.
22. Shurtleff SA, Buijs A, Behm FG, Rubnitz JE, Raimondi SC, Hancock ML, et al. TEL/AML1 fusion resulting from a cryptic t(12;21) is the most common genetic lesion in pediatric ALL and defines a subgroup of patients with an excellent prognosis. *Leukemia*. 1995 Dec;9(12):1985-9.
23. Paulsson K, Johansson B. High hyperdiploid childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer*. 2009 Aug;48(8):637-60.
24. Greaves MF, Wiemels J. Origins of chromosome translocations in childhood leukaemia. *Nat Rev Cancer*. 2003 Sep;3(9):639-49.
25. Pui CH, Relling MV, Downing JR. Acute lymphoblastic leukemia. *N Engl J Med*. 2004 Apr 8;350(15):1535-48.

26. Wiemels JL, Cazzaniga G, Daniotti M, Eden OB, Addison GM, Masera G, et al. Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet*. 1999 Oct 30;354(9189):1499-503.
27. McHale CM, Wiemels JL, Zhang L, Ma X, Buffler PA, Guo W, et al. Prenatal origin of TEL-AML1-positive acute lymphoblastic leukemia in children born in California. *Genes Chromosomes Cancer*. 2003 May;37(1):36-43.
28. Greaves MF. Aetiology of acute leukaemia. *Lancet*. 1997 Feb 1;349(9048):344-9.
29. Gale KB, Ford AM, Repp R, Borkhardt A, Keller C, Eden OB, et al. Backtracking leukemia to birth: identification of clonotypic gene fusion sequences in neonatal blood spots. *Proc Natl Acad Sci U S A*. 1997 Dec 9;94(25):13950-4.
30. Greaves MF, Maia AT, Wiemels JL, Ford AM. Leukemia in twins: lessons in natural history. *Blood*. 2003 Oct 1;102(7):2321-33.
31. Mori H, Colman SM, Xiao Z, Ford AM, Healy LE, Donaldson C, et al. Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc Natl Acad Sci U S A*. 2002 Jun 11;99(12):8242-7.
32. Tenen DG. Disruption of differentiation in human cancer: AML shows the way. *Nat Rev Cancer*. 2003 Feb;3(2):89-101.
33. Andreasson P, Schwaller J, Anastasiadou E, Aster J, Gilliland DG. The expression of ETV6/CBFA2 (TEL/AML1) is not sufficient for the transformation of hematopoietic cell lines in vitro or the induction of hematologic disease in vivo. *Cancer Genet Cytogenet*. 2001 Oct 15;130(2):93-104.
34. Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. *Nat Rev Cancer*. 2006 Mar;6(3):193-203.
35. Mullighan CG, Downing JR. Genome-wide profiling of genetic alterations in acute lymphoblastic leukemia: recent insights and future directions. *Leukemia*. 2009 Jul;23(7):1209-18.
36. Mullighan CG, Downing JR. Global genomic characterization of acute lymphoblastic leukemia. *Semin Hematol*. 2009 Jan;46(1):3-15.
37. Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*. 2007 Apr 12;446(7137):758-64.
38. Irving JA, Bloodworth L, Bown NP, Case MC, Hogarth LA, Hall AG. Loss of heterozygosity in childhood acute lymphoblastic leukemia detected by genome-wide microarray single nucleotide polymorphism analysis. *Cancer Res*. 2005 Apr 15;65(8):3053-8.

39. Lilljebjorn H, Heidenblad M, Nilsson B, Lassen C, Horvat A, Heldrup J, et al. Combined high-resolution array-based comparative genomic hybridization and expression profiling of ETV6/RUNX1-positive acute lymphoblastic leukemias reveal a high incidence of cryptic Xq duplications and identify several putative target genes within the commonly gained region. *Leukemia*. 2007 Oct;21(10):2137-44.
40. Dunwell T, Hesson L, Rauch TA, Wang L, Clark RE, Dalloil A, et al. A genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers. *Mol Cancer*. 2010;9:44.
41. Dunwell TL, Hesson LB, Pavlova T, Zabarovska V, Kashuba V, Catchpoole D, et al. Epigenetic analysis of childhood acute lymphoblastic leukemia. *Epigenetics*. 2009 Apr;4(3):185-93.
42. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*. 2002 Mar;1(2):133-43.
43. Horwitz M. The genetics of familial leukemia. *Leukemia*. 1997 Aug;11(8):1347-59.
44. Felix CA, D'Amico D, Mitsudomi T, Nau MM, Li FP, Fraumeni JF, Jr., et al. Absence of hereditary p53 mutations in 10 familial leukemia pedigrees. *J Clin Invest*. 1992 Aug;90(2):653-8.
45. Bridges JM, Nelson MG. Familial leukaemia. *Acta Haematol*. 1961;26:246-51.
46. Cramblett HG, Friedman JL, Najjar S. Leukemia in an infant born of a mother with leukemia. *N Engl J Med*. 1958 Oct 9;259(15):727-9.
47. Anderson RC. Familial leukemia; a report of leukemia in five siblings, with a brief review of the genetic aspects of this disease. *AMA Am J Dis Child*. 1951 Mar;81(3):313-22.
48. Gunz FW, Fitzgerald PH, Crossen PE, Mackenzie IS, Powles CP, Jensen GR. Multiple cases of leukemia in a sibship. *Blood*. 1966 Apr;27(4):482-9.
49. Schmiegelow K, Hjalgrim H. Is the risk of acute lymphoblastic leukemia reduced in siblings to children with the disease? A novel hypothesis explored by international collaboration. *Leukemia*. 2006 Jul;20(7):1206-8.
50. Winther JF, Sankila R, Boice JD, Tulinius H, Bautz A, Barlow L, et al. Cancer in siblings of children with cancer in the Nordic countries: a population-based cohort study. *Lancet*. 2001 Sep 1;358(9283):711-7.

51. Hawkins MM, Draper GJ, Winter DL. Cancer in the offspring of survivors of childhood leukaemia and non-Hodgkin lymphomas. *Br J Cancer*. 1995 Jun;71(6):1335-9.
52. Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads". *Am J Epidemiol*. 1998 Nov 1;148(9):893-901.
53. Kim AS, Eastmond DA, Preston RJ. Childhood acute lymphocytic leukemia and perspectives on risk assessment of early-life stage exposures. *Mutat Res*. 2006 Nov-Dec;613(2-3):138-60.
54. Shu XO, Stewart P, Wen WQ, Han D, Potter JD, Buckley JD, et al. Parental occupational exposure to hydrocarbons and risk of acute lymphocytic leukemia in offspring. *Cancer Epidemiol Biomarkers Prev*. 1999 Sep;8(9):783-91.
55. McKinney PA, Fear NT, Stockton D. Parental occupation at periconception: findings from the United Kingdom Childhood Cancer Study. *Occup Environ Med*. 2003 Dec;60(12):901-9.
56. Feychting M, Floderus B, Ahlbom A. Parental occupational exposure to magnetic fields and childhood cancer (Sweden). *Cancer Causes Control*. 2000 Feb;11(2):151-6.
57. Shu XO, Gao YT, Brinton LA, Linet MS, Tu JT, Zheng W, et al. A population-based case-control study of childhood leukemia in Shanghai. *Cancer*. 1988 Aug 1;62(3):635-44.
58. Shu XO, Reaman GH, Lampkin B, Sather HN, Pendergrass TW, Robison LL. Association of paternal diagnostic X-ray exposure with risk of infant leukemia. Investigators of the Childrens Cancer Group. *Cancer Epidemiol Biomarkers Prev*. 1994 Dec;3(8):645-53.
59. Shu XO, Ross JA, Pendergrass TW, Reaman GH, Lampkin B, Robison LL. Parental alcohol consumption, cigarette smoking, and risk of infant leukemia: a Childrens Cancer Group study. *J Natl Cancer Inst*. 1996 Jan 3;88(1):24-31.
60. Ji BT, Shu XO, Linet MS, Zheng W, Wacholder S, Gao YT, et al. Paternal cigarette smoking and the risk of childhood cancer among offspring of nonsmoking mothers. *J Natl Cancer Inst*. 1997 Feb 5;89(3):238-44.
61. Infante-Rivard C, Krajcinovic M, Labuda D, Sinnett D. Childhood acute lymphoblastic leukemia associated with parental alcohol consumption and polymorphisms of carcinogen-metabolizing genes. *Epidemiology*. 2002 May;13(3):277-81.

62. Wen W, Shu XO, Potter JD, Severson RK, Buckley JD, Reaman GH, et al. Parental medication use and risk of childhood acute lymphoblastic leukemia. *Cancer*. 2002 Oct 15;95(8):1786-94.
63. Ma X, Buffler PA, Gunier RB, Dahl G, Smith MT, Reinier K, et al. Critical windows of exposure to household pesticides and risk of childhood leukemia. *Environ Health Perspect*. 2002 Sep;110(9):955-60.
64. van Steensel-Moll HA, Valkenburg HA, Vandenbroucke JP, van Zanen GE. Are maternal fertility problems related to childhood leukaemia? *Int J Epidemiol*. 1985 Dec;14(4):555-9.
65. Jensen CD, Block G, Buffler P, Ma X, Selvin S, Month S. Maternal dietary risk factors in childhood acute lymphoblastic leukemia (United States). *Cancer Causes Control*. 2004 Aug;15(6):559-70.
66. Infante-Rivard C, Siemiatycki J, Lakhani R, Nadon L. Maternal exposure to occupational solvents and childhood leukemia. *Environ Health Perspect*. 2005 Jun;113(6):787-92.
67. van Steensel-Moll HA, Valkenburg HA, van Zanen GE. Childhood leukemia and parental occupation. A register-based case-control study. *Am J Epidemiol*. 1985 Feb;121(2):216-24.
68. Infante-Rivard C, Mur P, Armstrong B, Alvarez-Dardet C, Bolumar F. Acute lymphoblastic leukaemia among Spanish children and mothers' occupation: a case-control study. *J Epidemiol Community Health*. 1991 Mar;45(1):11-5.
69. Infante-Rivard C, Deadman JE. Maternal occupational exposure to extremely low frequency magnetic fields during pregnancy and childhood leukemia. *Epidemiology*. 2003 Jul;14(4):437-41.
70. Infante-Rivard C, Labuda D, Krajcinovic M, Sinnett D. Risk of childhood leukemia associated with exposure to pesticides and with gene polymorphisms. *Epidemiology*. 1999 Sep;10(5):481-7.
71. Freedman DM, Stewart P, Kleinerman RA, Wacholder S, Hatch EE, Tarone RE, et al. Household solvent exposures and childhood acute lymphoblastic leukemia. *Am J Public Health*. 2001 Apr;91(4):564-7.
72. Hatch EE, Linet MS, Kleinerman RA, Tarone RE, Severson RK, Hartsock CT, et al. Association between childhood acute lymphoblastic leukemia and use of electrical appliances during pregnancy and childhood. *Epidemiology*. 1998 May;9(3):234-45.
73. Mucci LA, Granath F, Cnattingius S. Maternal smoking and childhood leukemia and lymphoma risk among 1,440,542 Swedish children. *Cancer Epidemiol Biomarkers Prev*. 2004 Sep;13(9):1528-33.

74. Thompson JR, Gerald PF, Willoughby ML, Armstrong BK. Maternal folate supplementation in pregnancy and protection against acute lymphoblastic leukaemia in childhood: a case-control study. *Lancet*. 2001 Dec 8;358(9297):1935-40.
75. Ou SX, Han D, Severson RK, Chen Z, Neglia JP, Reaman GH, et al. Birth characteristics, maternal reproductive history, hormone use during pregnancy, and risk of childhood acute lymphocytic leukemia by immunophenotype (United States). *Cancer Causes Control*. 2002 Feb;13(1):15-25.
76. Shaw AK, Infante-Rivard C, Morrison HI. Use of medication during pregnancy and risk of childhood leukemia (Canada). *Cancer Causes Control*. 2004 Nov;15(9):931-7.
77. Infante-Rivard C, Fortier I, Olson E. Markers of infection, breast-feeding and childhood acute lymphoblastic leukaemia. *Br J Cancer*. 2000 Dec;83(11):1559-64.
78. Petridou E, Ntouvelis E, Dessypris N, Terzidis A, Trichopoulos D. Maternal diet and acute lymphoblastic leukemia in young children. *Cancer Epidemiol Biomarkers Prev*. 2005 Aug;14(8):1935-9.
79. Steffen C, Auclerc MF, Auvrignon A, Baruchel A, Kebaili K, Lambilliotte A, et al. Acute childhood leukaemia and environmental exposure to potential sources of benzene and other hydrocarbons; a case-control study. *Occup Environ Med*. 2004 Sep;61(9):773-8.
80. Noshchenko AG, Zamostyan PV, Bondar OY, Drozdova VD. Radiation-induced leukemia risk among those aged 0-20 at the time of the Chernobyl accident: a case-control study in the Ukraine. *Int J Cancer*. 2002 Jun 1;99(4):609-18.
81. Ahlbom A, Day N, Feychting M, Roman E, Skinner J, Dockerty J, et al. A pooled analysis of magnetic fields and childhood leukaemia. *Br J Cancer*. 2000 Sep;83(5):692-8.
82. Bianchi N, Crosignani P, Rovelli A, Tittarelli A, Carnelli CA, Rossitto F, et al. Overhead electricity power lines and childhood leukemia: a registry-based, case-control study. *Tumori*. 2000 May-Jun;86(3):195-8.
83. Draper G, Vincent T, Kroll ME, Swanson J. Childhood cancer in relation to distance from high voltage power lines in England and Wales: a case-control study. *BMJ*. 2005 Jun 4;330(7503):1290.
84. Cnattingius S, Zack MM, Ekblom A, Gunnarskog J, Kreuger A, Linet M, et al. Prenatal and neonatal risk factors for childhood lymphatic leukemia. *J Natl Cancer Inst*. 1995 Jun 21;87(12):908-14.

85. Infante-Rivard C, Mathonnet G, Sinnett D. Risk of childhood leukemia associated with diagnostic irradiation and polymorphisms in DNA repair genes. *Environ Health Perspect.* 2000 Jun;108(6):495-8.
86. Infante-Rivard C. Diagnostic x rays, DNA repair genes and childhood acute lymphoblastic leukemia. *Health Phys.* 2003 Jul;85(1):60-4.
87. Infante-Rivard C, Amre D, Sinnett D. GSTT1 and CYP2E1 polymorphisms and trihalomethanes in drinking water: effect on childhood leukemia. *Environ Health Perspect.* 2002 Jun;110(6):591-3.
88. Hjalgrim LL, Westergaard T, Rostgaard K, Schmiegelow K, Melbye M, Hjalgrim H, et al. Birth weight as a risk factor for childhood leukemia: a meta-analysis of 18 epidemiologic studies. *Am J Epidemiol.* 2003 Oct 15;158(8):724-35.
89. Yeazel MW, Ross JA, Buckley JD, Woods WG, Ruccione K, Robison LL. High birth weight and risk of specific childhood cancers: a report from the Children's Cancer Group. *J Pediatr.* 1997 Nov;131(5):671-7.
90. Ross JA, Potter JD, Shu XO, Reaman GH, Lampkin B, Robison LL. Evaluating the relationships among maternal reproductive history, birth characteristics, and infant leukemia: a report from the Children's Cancer Group. *Ann Epidemiol.* 1997 Apr;7(3):172-9.
91. Okcu MF, Goodman KJ, Carozza SE, Weiss NS, Burau KD, Bleyer WA, et al. Birth weight, ethnicity, and occurrence of cancer in children: a population-based, incident case-control study in the State of Texas, USA. *Cancer Causes Control.* 2002 Sep;13(7):595-602.
92. Paltiel O, Harlap S, Deutsch L, Knaanie A, Massalha S, Tiram E, et al. Birth weight and other risk factors for acute leukemia in the Jerusalem Perinatal Study cohort. *Cancer Epidemiol Biomarkers Prev.* 2004 Jun;13(6):1057-64.
93. McKinney PA, Juszczak E, Findlay E, Smith K, Thomson CS. Pre- and perinatal risk factors for childhood leukaemia and other malignancies: a Scottish case control study. *Br J Cancer.* 1999 Aug;80(11):1844-51.
94. Jourdan-Da Silva N, Perel Y, Mechinaud F, Plouvier E, Gandemer V, Lutz P, et al. Infectious diseases in the first year of life, perinatal characteristics and childhood acute leukaemia. *Br J Cancer.* 2004 Jan 12;90(1):139-45.
95. Ma X, Buffler PA, Wiemels JL, Selvin S, Metayer C, Loh M, et al. Ethnic difference in daycare attendance, early infections, and risk of childhood acute lymphoblastic leukemia. *Cancer Epidemiol Biomarkers Prev.* 2005 Aug;14(8):1928-34.

96. Chan LC, Lam TH, Li CK, Lau YL, Yuen HL, Lee CW, et al. Is the timing of exposure to infection a major determinant of acute lymphoblastic leukaemia in Hong Kong? *Paediatr Perinat Epidemiol*. 2002 Apr;16(2):154-65.
97. Schuz J, Kaletsch U, Meinert R, Kaatsch P, Michaelis J. Association of childhood leukaemia with factors related to the immune system. *Br J Cancer*. 1999 May;80(3-4):585-90.
98. Schuz J, Morgan G, Bohler E, Kaatsch P, Michaelis J. Atopic disease and childhood acute lymphoblastic leukemia. *Int J Cancer*. 2003 Jun 10;105(2):255-60.
99. Rosenbaum PF, Buck GM, Brecher ML. Allergy and infectious disease histories and the risk of childhood acute lymphoblastic leukaemia. *Paediatr Perinat Epidemiol*. 2005 Mar;19(2):152-64.
100. Groves FD, Gridley G, Wacholder S, Shu XO, Robison LL, Neglia JP, et al. Infant vaccinations and risk of childhood acute lymphoblastic leukaemia in the USA. *Br J Cancer*. 1999 Sep;81(1):175-8.
101. Ma X, Does MB, Metayer C, Russo C, Wong A, Buffler PA. Vaccination history and risk of childhood leukaemia. *Int J Epidemiol*. 2005 Oct;34(5):1100-9.
102. Nishi M, Miyake H. A case-control study of non-T cell acute lymphoblastic leukaemia of children in Hokkaido, Japan. *J Epidemiol Community Health*. 1989 Dec;43(4):352-5.
103. Murray L, McCarron P, Bailie K, Middleton R, Davey Smith G, Dempsey S, et al. Association of early life factors and acute lymphoblastic leukaemia in childhood: historical cohort study. *Br J Cancer*. 2002 Feb 1;86(3):356-61.
104. Shu XO, Linet MS, Steinbuch M, Wen WQ, Buckley JD, Neglia JP, et al. Breast-feeding and risk of childhood acute leukemia. *J Natl Cancer Inst*. 1999 Oct 20;91(20):1765-72.
105. Fradkin JE, Mills JL, Schonberger LB, Wysowski DK, Thomson R, Durako SJ, et al. Risk of leukemia after treatment with pituitary growth hormone. *JAMA*. 1993 Dec 15;270(23):2829-32.
106. Heath CW, Jr., Hasterlik RJ. Leukemia among children in a suburban community. 1963. *CA Cancer J Clin*. 1990 Jan-Feb;40(1):27-50.
107. Steinmaus C, Lu M, Todd RL, Smith AH. Probability estimates for the unique childhood leukemia cluster in Fallon, Nevada, and risks near other U.S. Military aviation facilities. *Environ Health Perspect*. 2004 May;112(6):766-71.
108. Heath CW, Jr. Community clusters of childhood leukemia and lymphoma: evidence of infection? *Am J Epidemiol*. 2005 Nov 1;162(9):817-22.

109. Kinlen L. Infections and immune factors in cancer: the role of epidemiology. *Oncogene*. 2004 Aug 23;23(38):6341-8.
110. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet*. 2009 Sep;41(9):1006-10.
111. Trevino LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet*. 2009 Sep;41(9):1001-5.
112. Franco RF, Simoes BP, Tone LG, Gabellini SM, Zago MA, Falcao RP. The methylenetetrahydrofolate reductase C677T gene polymorphism decreases the risk of childhood acute lymphocytic leukaemia. *Br J Haematol*. 2001 Dec;115(3):616-8.
113. Chatzidakis K, Goulas A, Athanassiadou-Piperopoulou F, Fidani L, Koliouskas D, Mirtsou V. Methylenetetrahydrofolate reductase C677T polymorphism: association with risk for childhood acute lymphoblastic leukemia and response during the initial phase of chemotherapy in greek patients. *Pediatr Blood Cancer*. 2006 Aug;47(2):147-51.
114. Krajcinovic M, Lamothe S, Labuda D, Lemieux-Blanchard E, Theoret Y, Moghrabi A, et al. Role of MTHFR genetic polymorphisms in the susceptibility to childhood acute lymphoblastic leukemia. *Blood*. 2004 Jan 1;103(1):252-7.
115. Gast A, Bermejo JL, Flohr T, Stanulla M, Burwinkel B, Schrappe M, et al. Folate metabolic gene polymorphisms and childhood acute lymphoblastic leukemia: a case-control study. *Leukemia*. 2007 Feb;21(2):320-5.
116. Taylor GM, Dearden S, Ravetto P, Ayres M, Watson P, Hussain A, et al. Genetic susceptibility to childhood common acute lymphoblastic leukaemia is associated with polymorphic peptide-binding pocket profiles in HLA-DPB1*0201. *Hum Mol Genet*. 2002 Jul 1;11(14):1585-97.
117. Taylor GM, Dearden S, Payne N, Ayres M, Gokhale DA, Birch JM, et al. Evidence that an HLA-DQA1-DQB1 haplotype influences susceptibility to childhood common acute lymphoblastic leukaemia in boys provides further support for an infection-related aetiology. *Br J Cancer*. 1998 Sep;78(5):561-5.
118. Krajcinovic M, Labuda D, Richer C, Karimi S, Sinnett D. Susceptibility to childhood acute lymphoblastic leukemia: influence of CYP1A1, CYP2D6, GSTM1, and GSTT1 genetic polymorphisms. *Blood*. 1999 Mar 1;93(5):1496-501.
119. Joseph T, Kusumakumary P, Chacko P, Abraham A, Radhakrishna Pillai M. Genetic polymorphism of CYP1A1, CYP2D6, GSTM1 and GSTT1 and susceptibility to acute lymphoblastic leukaemia in Indian children. *Pediatr Blood Cancer*. 2004 Oct;43(5):560-7.

120. Alves S, Amorim A, Ferreira F, Norton L, Prata MJ. The GSTM1 and GSTT1 genetic polymorphisms and susceptibility to acute lymphoblastic leukemia in children from north Portugal. *Leukemia*. 2002 Aug;16(8):1565-7.
121. Pakakasama S, Mukda E, Sasanakul W, Kadegasem P, Udomsubpayakul U, Thithapandha A, et al. Polymorphisms of drug-metabolizing enzymes and risk of childhood acute lymphoblastic leukemia. *Am J Hematol*. 2005 Jul;79(3):202-5.
122. Krajcinovic M, Labuda D, Sinnott D. Glutathione S-transferase P1 genetic polymorphisms and susceptibility to childhood acute lymphoblastic leukaemia. *Pharmacogenetics*. 2002 Nov;12(8):655-8.
123. Krajcinovic M, Sinnott H, Richer C, Labuda D, Sinnott D. Role of NQO1, MPO and CYP2E1 genetic polymorphisms in the susceptibility to childhood acute lymphoblastic leukemia. *Int J Cancer*. 2002 Jan 10;97(2):230-6.
124. Semsei AF, Erdelyi DJ, Ungvari I, Kamory E, Csokay B, Andrikovics H, et al. Association of some rare haplotypes and genotype combinations in the MDR1 gene with childhood acute lymphoblastic leukaemia. *Leuk Res*. 2008 Aug;32(8):1214-20.
125. Jamrozak K, Mlynarski W, Balcerczak E, Mistygacz M, Trelinska J, Mirowski M, et al. Functional C3435T polymorphism of MDR1 gene: an impact on genetic susceptibility and clinical outcome of childhood acute lymphoblastic leukemia. *Eur J Haematol*. 2004 May;72(5):314-21.
126. Krajcinovic M, Richer C, Sinnott H, Labuda D, Sinnott D. Genetic polymorphisms of N-acetyltransferases 1 and 2 and gene-gene interaction in the susceptibility to childhood acute lymphoblastic leukemia. *Cancer Epidemiol Biomarkers Prev*. 2000 Jun;9(6):557-62.
127. Eguchi-Ishimae M, Eguchi M, Ishii E, Knight D, Sadakane Y, Isoyama K, et al. The association of a distinctive allele of NAD(P)H:quinone oxidoreductase with pediatric acute lymphoblastic leukemias with MLL fusion genes in Japan. *Haematologica*. 2005 Nov;90(11):1511-5.
128. Joseph T, Kusumakumary P, Chacko P, Abraham A, Pillai MR. DNA repair gene XRCC1 polymorphisms in childhood acute lymphoblastic leukemia. *Cancer Lett*. 2005 Jan 10;217(1):17-24.
129. Infante-Rivard C, Krajcinovic M, Labuda D, Sinnott D. Parental smoking, CYP1A1 genetic polymorphisms and childhood leukemia (Quebec, Canada). *Cancer Causes Control*. 2000 Jul;11(6):547-53.
130. Morton NE. *Outline of Genetic Epidemiology*. Berlin: S. Karger; 1982.

131. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001 Feb 15;409(6822):860-921.
132. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001 Feb 16;291(5507):1304-51.
133. Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res*. 1998 Dec;8(12):1229-31.
134. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001 Jan 1;29(1):308-11.
135. Knight JC. Regulatory polymorphisms underlying complex disease traits. *J Mol Med*. 2005 Feb;83(2):97-109.
136. Pastinen T, Ge B, Hudson TJ. Influence of human genome polymorphism on gene expression. *Hum Mol Genet*. 2006 Apr 15;15 Spec No 1:R9-16.
137. Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, et al. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat Genet*. 2009 Nov;41(11):1216-22.
138. Chorley BN, Wang X, Campbell MR, Pittman GS, Nouredine MA, Bell DA. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. *Mutat Res*. 2008 Jul-Aug;659(1-2):147-57.
139. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. *Nat Genet*. 2004 Sep;36(9):949-51.
140. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, et al. Large-scale copy number polymorphism in the human genome. *Science*. 2004 Jul 23;305(5683):525-8.
141. Kuiper RP, Ligtenberg MJ, Hoogerbrugge N, Geurts van Kessel A. Germline copy number variation and cancer risk. *Curr Opin Genet Dev*. 2010 Apr 6.
142. Tost J. DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. *Mol Biotechnol*. 2010 Jan;44(1):71-81.
143. Cooper DN, Ball EV, Krawczak M. The human gene mutation database. *Nucleic Acids Res*. 1998 Jan 1;26(1):285-7.

144. Rebbeck TR. Inherited predisposition and breast cancer: modifiers of BRCA1/2-associated breast cancer risk. *Environ Mol Mutagen*. 2002;39(2-3):228-34.
145. Lander ES. The new genomics: global views of biology. *Science*. 1996 Oct 25;274(5287):536-9.
146. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996 Sep 13;273(5281):1516-7.
147. Muller-Myhsok B, Abel L. Genetic analysis of complex diseases. *Science*. 1997 Feb 28;275(5304):1328-9; author reply 9-30.
148. Scott WK, Pericak-Vance MA, Haines JL. Genetic analysis of complex diseases. *Science*. 1997 Feb 28;275(5304):1327; author reply 9-30.
149. Long AD, Grote MN, Langley CH. Genetic analysis of complex diseases. *Science*. 1997 Feb 28;275(5304):1328; author reply 9-30.
150. Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of Genetic Epidemiology*: Oxford University Press; 1993.
151. Eeles RA, Ponder BAJ, Easton DF, Horwich A, editors. *Genetic predisposition to cancer*: Chapman & Hall Medical; 1996.
152. Thomas DC. *Statistical Methods in Genetic Epidemiology*: Oxford University Press; 2004.
153. Peltomaki P, Aaltonen LA, Sistonen P, Pylkkanen L, Mecklin JP, Jarvinen H, et al. Genetic mapping of a locus predisposing to human colorectal cancer. *Science*. 1993 May 7;260(5109):810-2.
154. Lindblom A, Tannergard P, Werelius B, Nordenskjold M. Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nat Genet*. 1993 Nov;5(3):279-82.
155. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*. 1990 Dec 21;250(4988):1684-9.
156. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science*. 1994 Sep 30;265(5181):2088-90.
157. Hartl DL, Clark AG. *Principles of population genetics*. Third Edition: Sinauer Associates, Inc. Publishers; 1997.
158. A haplotype map of the human genome. *Nature*. 2005 Oct 27;437(7063):1299-320.

159. Clark AG. The role of haplotypes in candidate gene studies. *Genet Epidemiol.* 2004 Dec;27(4):321-33.
160. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet.* 2006 Oct;7(10):781-91.
161. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet.* 1993 Mar;52(3):506-16.
162. Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet.* 2006 May;7(5):385-94.
163. Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res.* 1998 Dec;8(12):1273-88.
164. Morton NE, Collins A. Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci U S A.* 1998 Sep 15;95(19):11389-93.
165. Kistner EO, Shi M, Weinberg CR. Using cases and parents to study multiplicative gene-by-environment interaction. *Am J Epidemiol.* 2009 Aug 1;170(3):393-400.
166. Schaid DJ. Case-parents design for gene-environment interaction. *Genet Epidemiol.* 1999;16(3):261-73.
167. Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA. Genetic association analysis using data from triads and unrelated subjects. *Am J Hum Genet.* 2005 Apr;76(4):592-608.
168. Nagelkerke NJ, Hoebee B, Teunis P, Kimman TG. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *Eur J Hum Genet.* 2004 Nov;12(11):964-70.
169. Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am J Hum Genet.* 2005 Oct;77(4):627-36.
170. Risch NJ. Searching for genetic determinants in the new millennium. *Nature.* 2000 Jun 15;405(6788):847-56.
171. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 2007 Oct 18;449(7164):851-61.
172. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. *Science.* 2005 Apr 15;308(5720):385-9.

173. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet.* 2007 May;39(5):631-7.
174. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature.* 2007 Jun 28;447(7148):1087-93.
175. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature.* 2008 Apr 3;452(7187):633-7.
176. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature.* 2008 Apr 3;452(7187):638-42.
177. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet.* 2008 May;40(5):616-22.
178. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet.* 2007 Aug;39(8):984-8.
179. Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet.* 2008 May;40(5):631-7.
180. Hardy J, Singleton A. Genomewide association studies and human disease. *N Engl J Med.* 2009 Apr 23;360(17):1759-68.
181. McPherson JD, Marra M, Hillier L, Waterston RH, Chinwalla A, Wallis J, et al. A physical map of the human genome. *Nature.* 2001 Feb 15;409(6822):934-41.
182. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009 Oct 8;461(7265):747-53.
183. Maher B. Personal genomes: The case of the missing heritability. *Nature.* 2008 Nov 6;456(7218):18-21.
184. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol.* 1980 Oct;112(4):467-70.
185. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet.* 2002 Oct 1;11(20):2463-8.

186. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered*. 2003;56(1-3):73-82.
187. Xu J, Lowey J, Wiklund F, Sun J, Lindmark F, Hsu FC, et al. The interaction of four genes in the inflammation pathway significantly predicts prostate cancer risk. *Cancer Epidemiol Biomarkers Prev*. 2005 Nov;14(11 Pt 1):2563-8.
188. Andrew AS, Nelson HH, Kelsey KT, Moore JH, Meng AC, Casella DP, et al. Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. *Carcinogenesis*. 2006 May;27(5):1030-7.
189. Manuguerra M, Matullo G, Veglia F, Autrup H, Dunning AM, Garte S, et al. Multi-factor dimensionality reduction applied to a large prospective investigation on gene-gene and gene-environment interactions. *Carcinogenesis*. 2007 Feb;28(2):414-22.
190. Siemiatycki J, Thomas DC. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int J Epidemiol*. 1981 Dec;10(4):383-7.
191. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays*. 2005 Jun;27(6):637-46.
192. Barker DJ. Fetal origins of coronary heart disease. *BMJ*. 1995 Jul 15;311(6998):171-4.
193. Barker DJ. The origins of the developmental origins theory. *J Intern Med*. 2007 May;261(5):412-7.
194. Mitchell LE, Weinberg CR. Evaluation of offspring and maternal genetic effects on disease risk using a family-based approach: the "pent" design. *Am J Epidemiol*. 2005 Oct 1;162(7):676-85.
195. Thomas DC. The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*. 2005 Mar;14(3):557-9.
196. Thomas DC, Baurley JW, Brown EE, Figueiredo JC, Goldstein A, Hazra A, et al. Approaches to complex pathways in molecular epidemiology: summary of a special conference of the American Association for Cancer Research. *Cancer Res*. 2008 Dec 15;68(24):10028-30.
197. Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P, et al. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiol Biomarkers Prev*. 2004 Jun;13(6):1013-21.

198. Hung RJ, Baragatti M, Thomas D, McKay J, Szeszenia-Dabrowska N, Zaridze D, et al. Inherited predisposition of lung cancer: a hierarchical modeling approach to DNA repair and cell cycle control pathways. *Cancer Epidemiol Biomarkers Prev*. 2007 Dec;16(12):2736-44.
199. Greenland S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Stat Med*. 1992 Jan 30;11(2):219-30.
200. Witte JS, Greenland S, Haile RW, Bird CL. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology*. 1994 Nov;5(6):612-21.
201. Conti DV, Cortessis V, Molitor J, Thomas DC. Bayesian modeling of complex metabolic pathways. *Hum Hered*. 2003;56(1-3):83-93.
202. Thomas D. Gene-environment-wide association studies: emerging approaches. *Nat Rev Genet*. 2010 Mar 9;11(4):259-72.
203. Golias CH, Charalabopoulos A, Charalabopoulos K. Cell proliferation and cell cycle control: a mini review. *Int J Clin Pract*. 2004 Dec;58(12):1134-41.
204. Malumbres M, Barbacid M. To cycle or not to cycle: a critical decision in cancer. *Nat Rev Cancer*. 2001 Dec;1(3):222-31.
205. Sherr CJ. The Pezcoller lecture: cancer cell cycles revisited. *Cancer Res*. 2000 Jul 15;60(14):3689-95.
206. Kastan MB, Bartek J. Cell-cycle checkpoints and cancer. *Nature*. 2004 Nov 18;432(7015):316-23.
207. Spry M, Scott T, Pierce H, D'Orazio JA. DNA repair pathways and hereditary cancer susceptibility syndromes. *Front Biosci*. 2007;12:4191-207.
208. Heinen CD, Schmutte C, Fishel R. DNA repair and tumorigenesis: lessons from hereditary cancer syndromes. *Cancer Biol Ther*. 2002 Sep-Oct;1(5):477-85.
209. Li C, Wang LE, Wei Q. DNA repair phenotype and cancer susceptibility--a mini review. *Int J Cancer*. 2009 Mar 1;124(5):999-1007.
210. Goode EL, Ulrich CM, Potter JD. Polymorphisms in DNA repair genes and associations with cancer risk. *Cancer Epidemiol Biomarkers Prev*. 2002 Dec;11(12):1513-30.
211. McKinnon PJ, Caldecott KW. DNA strand break repair and human genetic disease. *Annu Rev Genomics Hum Genet*. 2007;8:37-55.

212. Fortini P, Dogliotti E. Base damage and single-strand break repair: mechanisms and functional significance of short- and long-patch repair subpathways. *DNA Repair (Amst)*. 2007 Apr 1;6(4):398-409.
213. Wood RD. Nucleotide excision repair in mammalian cells. *J Biol Chem*. 1997 Sep 19;272(38):23465-8.
214. Mostoslavsky R. DNA repair, insulin signaling and sirtuins: at the crossroads between cancer and aging. *Front Biosci*. 2008;13:6966-90.
215. Hsieh P, Yamane K. DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mech Ageing Dev*. 2008 Jul-Aug;129(7-8):391-407.

CHAPTER TWO

*INDIVIDUAL MAIN EFFECTS OF CANDIDATE
GENES ON THE SUSCEPTIBILITY TO
CHILDHOOD ACUTE LYMPHOBLASTIC LEUKEMIA*

CANDIDATE GENE ASSOCIATION STUDIES

My first objective was to investigate the putative main effects of selected gene variants on childhood ALL risk. A two-tier study design was used in which a population-based case-control study, conducted in 321 patients with pre-B ALL and 329 healthy controls, was performed in conjunction with a family-based analysis of 203 case-parent trios, to test for association. It should be noted however that cases used in the family-based association study were of the 321 used in the case-control analysis therefore these two analyses were not independent (Figure 1).

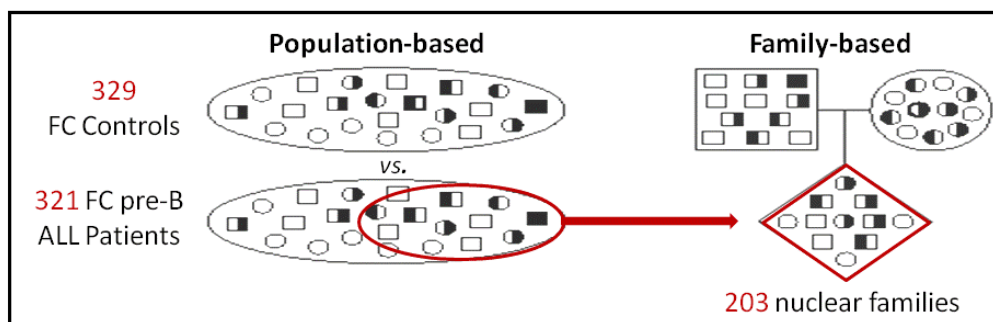


Figure 1. Design of the candidate gene association study

A two-tier design was used in which 321 pre-B ALL cases and 329 healthy controls were used in a population-based case-control association study, and of these patients, 203 had additional parental data available and were enrolled in the family-based association study. All individuals were of French-Canadian origin from the province of Quebec.

Individuals participating in the study were from the Quebec childhood ALL (QcALL) cohort and were enrolled in the study at the Sainte-Justine Hospital, Montreal between 1985 and 2006. Patients were diagnosed and/or treated in the Division of Hematology-Oncology of the Sainte-Justine Hospital, Montreal and included 190 males and 131 females between the ages of 0 and 18 years with a median age of 4.3 years. Healthy controls included 183 males and 146 females and were also recruited at the Sainte-Justin Hospital. The QcALL cohort consists of incident ALL cases, for which both normal and tumoral bio-specimen are available, of case-parents as well as unrelated controls. Detailed demographic, clinical and histopathological data are also available for most of the patients (Table 1), providing a unique opportunity for the genetic characterization of childhood ALL and its subtypes. Only pre-B ALL cases (no T-cell ALL) were investigated because of the higher prevalence of this subtype and also to decrease clinical heterogeneity. All study subjects were of French-Canadian origin residing in the province of Quebec, Canada as judged by their names, languages and places of birth. The French-Canadian population is a relatively homogenous population due to particular demographic and historic characteristics. The virtues for using founder populations, such as the French-Canadians, for genetic epidemiology studies include increased genetic and environmental homogeneity, potentially simpler genetic architectures for complex diseases, and longer stretches of linkage disequilibrium facilitating association testing. In recruiting only French-Canadian individuals diagnosed and/or treated at the Sainte-Justine Hospital, we reduced ethnic admixture and regional variability in the frequency of particular ethnic backgrounds.

Table 1. Characteristics of the pre-B ALL patients from the Quebec Childhood ALL cohort

Patient characteristics	Cases, n (%)
Total number of patients*	321
Gender	
Male	190 (59.2)
Female	131 (40.8)
Age group, years	
≤ 1	8 (2.5)
1-10	235 (73.2)
> 10	53 (16.5)
N/D	25 (7.8)
Hyperdiploidy	
Positive [†]	113 (35.2)
Negative	167 (52.0)
N/D	41 (12.8)
Chromosomal translocations	
Absence of translocation	97 (30.2)
t(12;21) [†]	35 (10.9)
Other	12 (3.7)
N/D	177 (55.1)
Normal ploidy and absence of chromosomal translocations	43 (13.4)

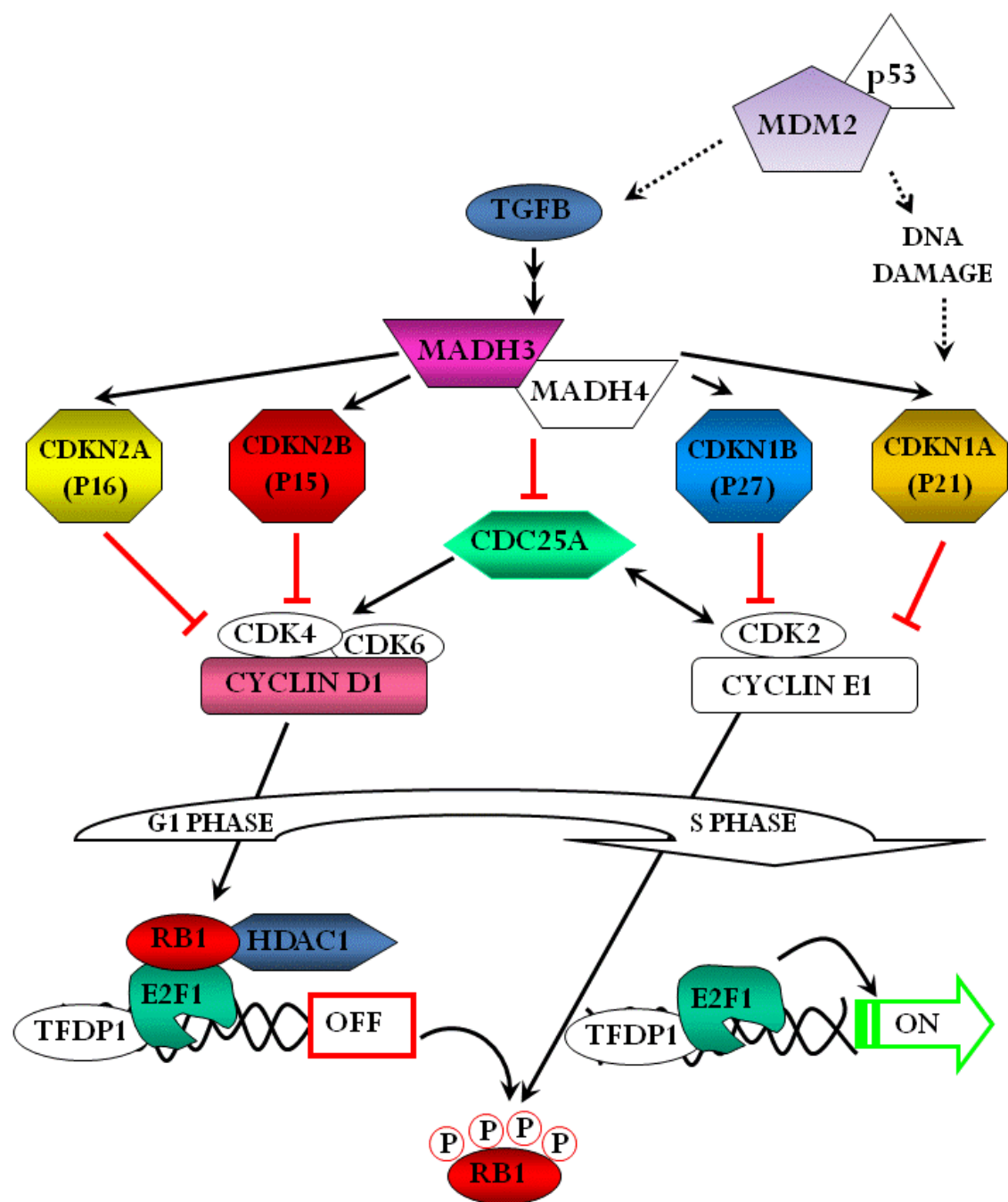
* Numbers are shown only for patients used in the candidate gene association studies presented here.

† Hyperdiploidy and the t(12;21) translocation are virtually mutually exclusive; only a single patient in our cohort carries both.

N/D indicates data not available.

Candidate genes were selected based on their key function in two cancer-related pathways: G1/S checkpoint regulation and double-strand break repair. The 2kb proximal promoter regions of the candidate genes were screened for polymorphisms and only common promoter SNPs (MAF \geq 5% in at least one of the case or control groups) were kept for investigation. In total 26 promoter SNPs from 12 candidate cell cycle genes and 20 promoter SNPs from 7 double-strand break repair genes were selected for analysis (Figure 2 and Table 2). Genotyping was performed using either the PCR-based allele-specific oligonucleotide (ASO) hybridization approach, or the allele-specific primer extension approach using the Luminex system. A 40 base pair deletion in the promoter region of the *MDM2* gene was also included in the analysis and was genotyped using gel electrophoresis to separate the length polymorphism PCR products. Hardy-Weinberg equilibrium (HWE) was tested using the X^2 goodness of fit test and the PedCheck Software was used to identify genotype incompatibilities using the familial data.

A)



B)

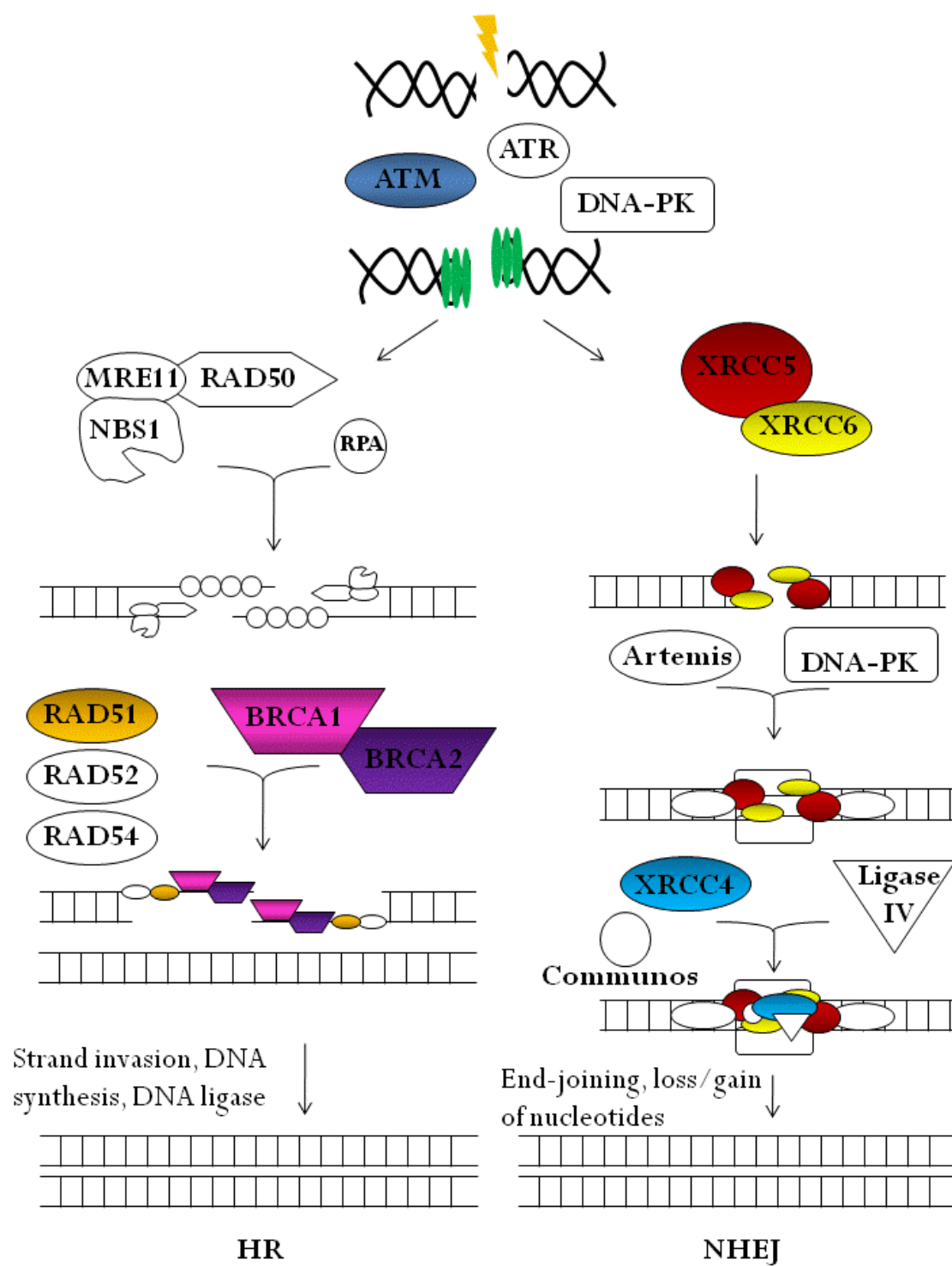


Figure 2. Candidate genes selected based on their function in two cancer-related pathways

This genetic association study focused on two candidate pathways **A)** The G1/S cell cycle checkpoint and **B)** DNA double-strand break damage repair (both homologous repair, HR, and non-homologous end-joining, NHEJ) pathways. A total of 26 promoter SNPs from 12 cell cycle genes and 20 promoter SNPs from 7 DNA repair genes (shown in colour in the respective diagrams) were genotyped among pre-B ALL cases, their parents and unrelated controls of the QcALL cohort. See Table 2 for details.

Table 2. Characteristics of genes and corresponding DNA variants genotyped in the association studies

Pathway, Gene (Chr.), DNA variant	dbSP	Position	MAF (%)
G1/S cell cycle checkpoint			
<i>CCND1</i> (11q13)			
-1938T>C	rs1944129	69,453,935	48.8
-1537INS C	rs36225395	69,454,336	45.2
<i>CDC25A</i> (3p21)			
-2030G>T	rs1903061	48,231,919	10.3
<i>CDKN1A</i> (6p21.2)			
-1284T>C	rs733590	36,645,203	36.2
-899T>G	rs762624	36,645,588	27.1
-791T>C	rs2395655	36,645,696	39.74
<i>CDKN1B</i> (12p13)			
-1857C>T	rs3759217	12,868,452	11.6
-1608G>A	rs35756741	12,868,701	8.6
-373G>T	rs36228499	12,869,936	43.4
<i>CDKN2A</i> (9p21)			
-222T>A	rs36228834	21,975,319	3.3
<i>CDKN2B</i> (9p21)			
-1270C>T	rs36229158	22,010,681	2.8
-593A>T,C	rs2069416	22,010,004	37.4/2.7
-287G>C	rs2069418	22,009,698	42.7
<i>E2F1</i> (20q11.2)			
-187C>T	rs3213141	32,274,380	24.1
<i>HDAC1</i> (1p34.1)			
-1269T>C	rs1741981	32,756,439	33.0
-455T>C	rs36212119	32,757,253	8.4
<i>MDM2</i> (12q14.3-q15)			
-1494A>G	rs1144944	69,200,485	49.2
-1174DEL AAAAAGC(40bp)	rs3730485	69,200,806-69,200,845	40.5
-182C>G	rs937282	69,201,797	48.3
+309T>G	rs2279744	69,202,580	36.6
<i>SMAD3</i> (15q21-q22)			
-1938T>C	rs36221701	67,356,489	12.8
<i>RB1</i> (13q14.1-q14.2)			
-1554C>A	rs1573601	48,876,357	24.8
<i>TGFB1</i> (19q13.1)			
-1886A>G	rs2317130	41,861,674	31.4
-1571G>A	rs4803457	41,861,359	39.4
-1550DEL AGG	rs11466313	41,861,338-41,861,337	31.0
-508G>A	rs1800469	41,860,296	31.3
DNA double-strand break repair			
<i>ATM</i> (11q22.3)			
-1206G>T	rs4987876	108,092,637	9.6
-635T>A	rs228589	108,093,208	41.8

<i>BRCA1</i> (17q21)			
-1890T>C	rs4793204	41,279,298	32.4
-708A>G	rs799906	41,278,116	33.4
-598INS ACA	rs8176071	41,278,006-41,278,005	32.3
-588A>G	rs3092986	41,277,996	9.0
<i>BRCA2</i> (13q12.3)			
-1555C>A	rs206114	32,888,062	41.3
-1260DEL GTCTAA	rs3072036	32,888,357-32,888,362	39.6
-1144A>G	rs206115	32,888,473	40.9
-1134C>T	rs206116	32,888,483	40.2
-908C>T	rs206117	32,888,709	39.7
-254G>A	rs3092989	32,889,363	20.5
<i>RAD51</i> (15q15.1)			
-1185A>T	rs2619679	40,986,237	49.1
<i>XRCC4</i> (5q13-q14)			
-1864T>C	rs3763063	82,371,453	49.5
-1407C>G	rs1993947	82,371,910	8.4
<i>XRCC5</i> (2q35)			
-1379G>T	rs828907	216,972,732	47.1
-297C>T	rs11685387	216,973,814	25.6
<i>XRCC6</i> (22q11-q13)			
-1469C>T	rs28384701	42,016,526	4.5
-1296C>G	rs2267437	42,016,699	38.6
-731G>A	rs132770	42,017,264	25.1

DNA variant positions were numbered with respect to the first nucleotide of the first exon as +1, and the nucleotide immediately upstream as -1. Mapping of SNP positions is based on dbSNP build 130 and the GRCh37 human assembly (UCSC Genome Browser). Minor allele frequency (MAF) was calculated on a control cohort consisting of 329 healthy French-Canadian individuals from the QcALL cohort; overall frequencies are comparable to those reported for other populations of European descent.

Using both the case-control and case-parent trio data, we screened for associations with childhood ALL (allelic and genotypic). When applicable, haplotype-based association studies were performed to further capture the effects of *cis*-acting genetic variation in the promoter regions of these genes. Traditional statistical methods including chi-square or Fisher's exact tests (2-sided) and unconditional logistic regression were used to compare allele/genotype/haplotype carriership and measure association in the case-control design, and the family-based association test (FBAT), was used to test for deviations from random Mendelian transmission of alleles/haplotypes to affected offspring among case-parent trios. Together, these methods allowed efficient use of data from both unrelated individuals and available parents. Our sample size of 321 cases and roughly the same number of unrelated controls provided 80% power at the 5% level to detect a minimum odds ratio (OR) of 1.9 with minor allele frequencies $MAF \geq 0.05$ and of 1.5 with $MAF \geq 0.15$ (Figure 3), while our family-based design consisting of 203 case-parent trios yielded 80% power at the 5% level to detect a minimum OR of 1.65 with $MAF \geq 0.15$ (Figure 4). Note that a overlapping group of participants was used in both sets of power calculations, and that the case-control and family-based association tests were not independent and were used as complementary strategies in the identification of disease associations. See Appendix II for detailed power calculations.

Functional validation of the regulatory polymorphisms was performed in collaboration with colleagues in the laboratory. Both gene reporter and electrophoretic mobility shift assays allowed us to identify allele-specific differential binding and/or expression providing some insight into the putative function of the genetic variants in disease.

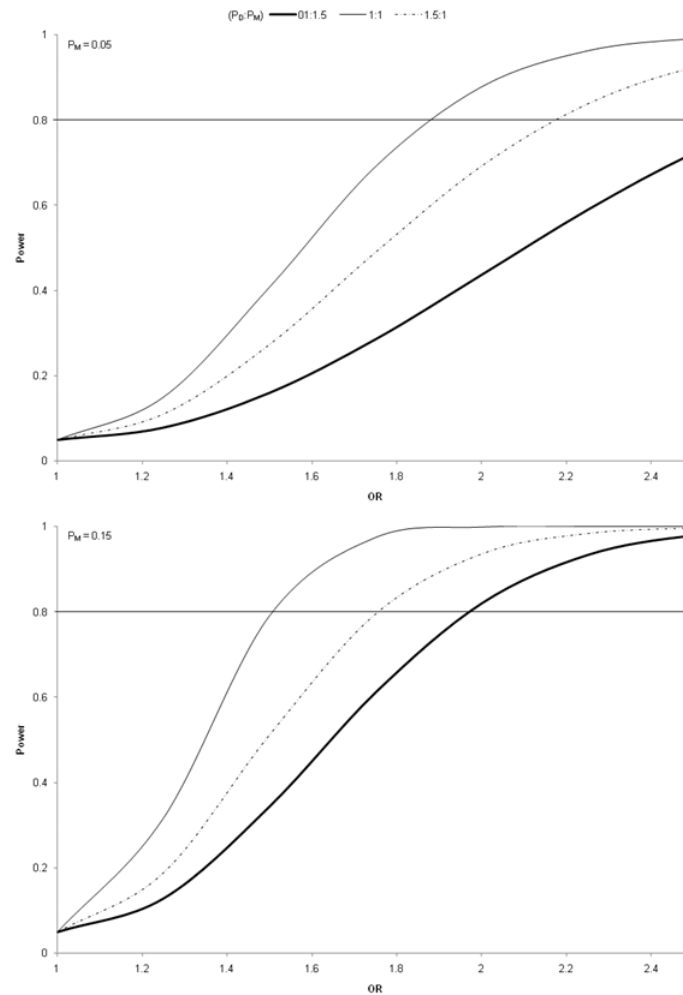


Figure 3. Power calculations for main effects using a case-control design

Power calculations when the measured locus M occurs with frequency P_M and the ratio of frequencies between M and the disease locus D is given by $P_D : P_M$. The measure of linkage disequilibrium between M and P was set at $D' = 0.8$. P_M was set at 5% and 15%. P_D varied from 1.5-times lower to 1.5-times higher than the marker allele frequency. A type I error rate of 0.05 (two-sided) was used for a sample size of $N=321$ and an unmatched case-control ratio of 1:1, assuming an overall disease prevalence rate of 0.0001. A multiplicative (log-additive) inheritance model was used and effect sizes (ORs) were allowed to vary from 1.0 to 3.0. Horizontal reference line indicates 80% power. See Appendix II for detailed power calculations.

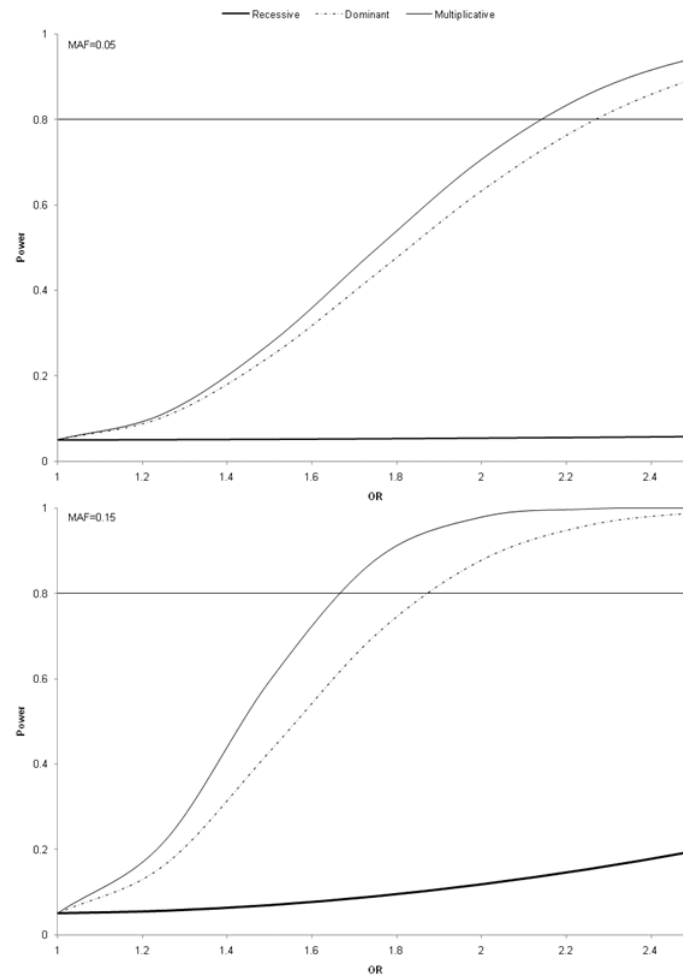


Figure 4. Power calculations for main effects using a family-based case-parent trio design

Recessive, dominant and multiplicative inheritance models were tested for minor allele frequencies (MAFs) of 0.05 and 0.15 and for main effect sizes (ORs) varying from 1.0 to 3.0. A type I error rate of 0.05 (two-sided) was used for a sample size consisting of 203 complete trios (father, mother and one affected child) and an overall disease prevalence rate of 0.0001. Horizontal reference line indicates 80% power. See Appendix II for detailed power calculations.

These initial studies lead to the publication of three articles which are presented in this chapter:

- Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia (Healy et al. (2007) *Blood* 109(2):683-92);
- No evidence for association between *TGFB1* promoter SNPs and the risk of childhood pre-B acute lymphoblastic leukemia among French Canadians (Healy et al. (2009) *Haematologica* 94(7):1034-5);
- Functional impact of sequence variation in the promoter region of *TGFB1* (Healy et al. (2009) *International Journal of Cancer* 125(6):1483-9).

Tables and figures presenting the complete results of the association studies for all candidate genes/variants are available in Appendix II.

GENOME-WIDE ASSOCIATION STUDIES

In addition to this candidate gene association study, we have joined forces with a large group of international collaborators from Europe, Asia and the United States to form the International Acute Lymphoblastic Leukemia Genetics Consortium (IALLGC) whose aim is to identify and characterize low-penetrance susceptibility variants for childhood ALL through large-scale association-based analyses.

Two independent genome-wide association studies recently provided convincing evidence that common, inherited genetic variation contributes to childhood ALL susceptibility. Both studies identified overlapping variants in genes involved in transcriptional regulation and differentiation of B-cell progenitors that were associated with increased risk of childhood B-cell ALL. One of the immediate goals of the consortium is to provide opportunities for replication of genetic variants shown to be associated with childhood ALL. To this extent, we attempted to validate 15 of the initial GWAS signals from both studies in the QcALL cohort and successfully replicated the association of 5 SNPs within the *ARID5B* gene. This publication is also presented in this chapter:

- Replication analysis confirms the association of *ARID5B* with childhood B-cell acute lymphoblastic leukemia (Healy et al. (2010) *Haematologica*, Epub ahead of print, doi:10.3324/haematol.2010.022459).

Promoter SNPs in G1/S Checkpoint Regulators and their Impact on the Susceptibility to Childhood Leukemia

Jasmine Healy,¹ Hélène Bélanger,¹ Patrick Beaulieu,¹ Mathieu Larivière,¹
Damian Labuda,^{1,2} and Daniel Sinnett^{1,2}

¹Division of Hematology-Oncology, Sainte-Justine Hospital Research Center, University of Montreal, QC, Canada;

²Department of Pediatrics, Faculty of Medicine, University of Montreal, QC, Canada

This research was originally published in *Blood*. Healy J, Bélanger H, Beaulieu P, Larivière M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood*. 2007 Jan; 109(2): 683-692. © the American Society of Hematology. The article is included with the permission of the authors and the journal.

Author Contributions

For this article, I carried out 80% of the work including the genotyping and all of the statistical analyses, and I wrote the manuscript. H. Bélanger was involved in pSNP discovery; M. Larivière carried out the EMSAs; P. Beaulieu performed *in silico* studies. D. Labuda was involved in the world population data collection for pSNP discovery. D. Sinnett is principle investigator and participated in the design and coordination of the study as well as writing the manuscript. All authors read and approved the final manuscript.

Abstract

Mutations leading to alteration of cell cycle checkpoint functions are a common feature of most cancers. Because of the highly regulated nature of the cell cycle, it seems likely that variation in gene dosage of key components due to functional regulatory polymorphisms could play an important role in cancer development. Here we provide evidence of the involvement of promoter SNPs (pSNPs) in cyclin-dependent kinase inhibitor genes *CDKN2A*, *CDKN2B*, *CDKN1A* and *CDKN1B*, in the etiology of childhood pre-B acute lymphoblastic leukemia (ALL). A case-control study conducted in 240 pre-B ALL patients and 277 healthy controls combined with a family-based analysis using 135 parental trios, all of French-Canadian origin, were used to evaluate single site genotypic as well as multilocus haplotypic associations for a total of ten pSNPs. Using both study designs, we showed evidence of association between variants *CDKN2A* -222A, *CDKN2B* -593A and *CDKN1B* -1608A and an increased risk of ALL. These findings suggest that variable expression levels of cell cycle inhibitor genes *CDKN2A*, *CDKN2B* and *CDKN1B* due to regulatory polymorphisms could indeed influence the risk of childhood pre-B ALL and contribute to carcinogenesis.

Introduction

Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer. The etiology of this hematologic malignancy might be explained by a combination of genetic susceptibility and environmental exposure during early development in fetal life and infancy. Assuming that genes modulate individual responses to exogenous and/or endogenous factors, they would thereby also influence an individual's risk of cancer (1). Consistent with this paradigm, it has been shown that childhood leukemogenesis is associated with genetic variability in xenobiotic metabolism (2-10), oxidative stress response (9, 11, 12) and DNA repair (13, 14) pathways. However, little is known about the impact of genetic polymorphisms in cell cycle components, despite the fact that the cell cycle is a highly orchestrated biological process frequently altered in human cancers (15). A critical point in the cell cycle is the G1/S transition checkpoint, during which the cell is irreversibly committed to a new round of division (16). The cyclin-dependent kinase inhibitors (CDKIs) CDKN2A (p16^{INK4A}), CDKN2B (p16^{INK4B}), CDKN1A (p21^{Cip1/Waf1}), and CDKN1B (p27^{kip1}) are key regulators of the G1/S checkpoint, their concerted action preventing cells from undergoing subsequent division in response to oncogenic signaling or DNA damage (17). Accordingly, changes in their expression and/or activity due to polymorphisms might modify the susceptibility to cancer. Correlations between DNA variants in *CDKN2A*, *CDKN2B*, *CDKN1A*, and *CDKN1B* and cancer susceptibility have been reported, particularly in breast, prostate, and skin carcinomas (18-24). In addition, single-nucleotide polymorphisms (SNPs) in gene promoter sequences (pSNPs) have recently gained much importance because of their quantitative impact on gene expression (25, 26). Several studies have suggested that because protein levels regulate many biological pathways, including the cell cycle, pSNPs could influence the overall outcome of these biological processes and thereby modify disease risk (27-33). In this report we performed an association study using both a case-control and family-based design in order to

assess the impact of proximal promoter SNPs in *CDKN2A*, *CDKN2B*, *CDKN1A*, and *CDKN1B* on the susceptibility to childhood pre-B ALL.

Patients, Materials, and Methods

Study population

The population under study and the inclusion criteria were described previously (6). Incident patients with childhood pre-B ALL ($n = 240$) were diagnosed in the Division of Hematology-Oncology of the Sainte-Justine Hospital in Montreal (QC, Canada), between October 1985 and November 2003. They comprised 141 boys and 99 girls with a median age of 4.6 years (range, 5 months to 18 years), all of French-Canadian descent, from the province of Quebec. A number of these patients (135), from whom parent DNA was available, were also enrolled in the family-based study. The healthy controls ($n = 277$) consisted of French-Canadian volunteers recruited while using clinical departments other than Hematology-Oncology of the Sainte-Justine Hospital. The CHU Sainte-Justine Institutional Review Board approved the research protocol, and informed consent was obtained from all participating individuals and/or their parents.

Genotyping of promoter SNPs

DNA was isolated from either buccal epithelial cells, peripheral blood, or bone marrow in remission as described previously (34). DNA segments containing the polymorphic sites were amplified by polymerase chain reaction (PCR) using a “touchdown” thermal cycling protocol (35). The resulting PCR products were dot-blotted in duplicate on a nylon membrane and assayed for the presence or absence of variants by hybridization in parallel with allele-specific

oligonucleotides (ASOs) as described in Labuda et al (36). The amplimers and oligonucleotide probes used for ASO analysis are given in Table 1.

EMSAs

Double-stranded oligonucleotide probes corresponding to the sequences surrounding the polymorphic sites were radiolabeled and allowed to interact with nuclear extracts prepared from HepG2 (hepatoma), Jeg-3 (choriocarcinoma), and HeLa (cervical carcinoma) cells as described in Belanger et al (37). Briefly, protein was quantified with the Bradford protein assay (Bio-Rad, Mississauga, ON, Canada). Nuclear extracts (5 µg) were incubated with 35 fmol radiolabeled double-stranded DNA probes and a buffer containing 50 mM Tris-HCl (pH 7.5), 5 mM MgCl₂, 2.5 mM EDTA, 2.5 mM DTT, 250 mM NaCl, 0.25 µg/µL poly deoxyinosinate-deoxycytidylate, and 20% glycerol, in a total volume of 10 µL for 20 minutes at room temperature. Complexes were separated in a 6% nondenaturing polyacrylamide gel (acrylamide-bisacrylamide, 60:1) in 0.5X Tris-borate-EDTA buffer (190 V at 4°C). For competition of binding, a 50-fold molar excess of competitor (either the unlabeled probe oligo nucleotide or the corresponding mutant oligo) was included. The oligonucleotides used in the electrophoretic mobility shift assays (EMSAs) are as follows (only the 5'-3' top strand of the double-stranded oligonucleotides are shown):

CDKN2A –222T>A, ACAACCTTCC(T/A)AACTGCCAAATTGAATCGGGGTGT;
CDKN2B –1270C>T,
 AATGCTACCCGGTTCCCTT(C/T)CCTGTCCAGGTGGATTT, –593A>T, C,
 GGATCTCAGATTCTTA(A/T/C)AGTATAATTTTTTTT, and –287C>G
 ATCTTAAGAAA(C/G)ACGGAGTTATTTTGA;
CDKN1A –1284T>C, TTCTGTTTTT(T/C)AGTGGGATTT, –899T>G,
 TGGGGAAAC(T/G)GGGGCTC, and –791T>C,
 ACAGAAGAAA(T/C)CCCTGTGGTT;
 and *CDKN1B* –1857C>T, TACCACAGGCTCAAGA(C/T)AGCTGCATTAA,

–1608G>A, GGTTTCCTGTCCAGG(G/A)ACATGCA, and –373G>T, TAAGCCCCGACCTCCCTCCCGCTCCTC(G/T)CCCGGGAAGCCGGGAC.

Statistical analysis

Hardy-Weinberg equilibrium was examined using the χ^2 test for goodness of fit. Fisher exact tests (2-sided) were used to compare allele/genotype/haplotype carriership in patients and controls. Crude odds ratios (ORs) are given with 95% confidence intervals (CIs). All analyses were carried out using STATA statistical software (release 9.1; StataCorp, College Station, TX). *CDKN2B*, *CDKN1A*, and *CDKN1B* haplotypes and corresponding frequencies were estimated using the PHASE software (version 2; University of Washington, Seattle) (38). Linkage disequilibrium between SNPs was tested with the Arlequin linkage utility software (version 2.00; University of Geneva, Switzerland) (39). An omnibus χ^2 test, implemented in the Evolutionary-Based Haplotype Analysis Package (eHap version 2.0; Carnegie Mellon University and University of Pittsburgh, PA) (40) was used to examine overall haplotype associations with disease phenotype. Transmission disequilibrium from parents to children of individual SNPs and corresponding haplotypes was assessed with FBAT (family-based association test) software (version 1.5.1; Program for Population Genetics, Harvard School of Public Health, Boston, MA) (41). A multiallelic test was also carried out in FBAT to obtain the global haplotype association significance in the family-based setting. Correction for multiple testing errors was performed using the false discovery rate (FDR) principle (42), with a predetermined type I error rate set at 10%.

Results

Promoter SNPs

The detection of proximal promoter variants in a population panel consisting of 40 unrelated individuals (8 Africans, 8 Europeans, 8 Asians, 8 Middle-Easterners, and 8 Amerindians) was performed by PCR-based denaturing high-performance liquid chromatography (dHPLC) analysis followed by direct sequencing as described in Sinnott et al. (43). The targeted promoter region was arbitrarily defined as the 2-kb sequence upstream of the transcriptional initiation site. A total of 40 sequence variants were identified for the genes *CDKN2A*, *CDKN2B*, *CDKN1A*, and *CDKN1B*, including 21 that were previously reported in public databases (Figure 1). For the purpose of this study targeting French-Canadians, we considered only pSNPs that were common among Europeans; in other words, those that were found at least twice among the 8 European patients initially screened. This led to the genotyping of 10 pSNPs: *CDKN2A*, -222T>A; *CDKN2B*, -1270C>T, -593A>T,C and -287C>G; *CDKN1A*, -1284T>C, -899T>G, and -791T>C; and *CDKN1B*, -1857C>T, -1608G>A, and -373G>T (Table 1). Of note, polymorphism -222T>A has been detected in the promoter region of *CDKN2A* in previous studies (19, 44, 45). The observed allele and genotype frequencies in children with ALL and in healthy controls are reported in Tables 2 and 3. In controls, the frequencies of 6 of these pSNPs were in agreement with those reported in other populations of European descent as per the National Center of Biotechnology Information (NCBI) database of SNPs (dbSNP) (46) and the International Haplotype Mapping (HapMap) Project (47) databases. All distributions were in Hardy-Weinberg equilibrium.

Predicted functional impact of the pSNPs

The screening of the promoter region for predicted transcription factor–binding sites (TFBSs) using matInspector (<http://www.genomatix.de/products/index.html>) (48) led to the identification of pSNPs that might create and/or disrupt some of these TFBSs (Table 4). The putative impact of these pSNPs on DNA-protein–binding capacity was further validated by EMSAs. In the latter, double stranded oligonucleotide probes corresponding to the sequences surrounding the polymorphic sites (“Patients, materials, and methods”) were radiolabeled and allowed to interact with nuclear extracts prepared from HepG2, Jeg-3, and HeLa cells, and differential allelic shifts were assessed (see Figure 2 for representative data). As indicated in Table 4, 7 of the tested pSNPs showed differential allelic shifts in at least 1 of the cell lines tested.

Single-locus analysis

First we assessed the involvement of the selected 10 pSNPs in childhood ALL by performing a case-control study. The estimated ORs and 95% CIs for the corresponding alleles and genotypes are given in Tables 2 and 3. The *CDKN2A* –222A allele was overrepresented in patients when compared with controls (7.5% versus 3.6%), as was the heterozygous –222TA genotype (13.2% versus 6.5%). Evidence of an increased risk of ALL among carriers of the *CDKN2A* –222A variant was demonstrated (OR = 2.2, 95% CI, 1.2-4.0, $P = .008$) and remained significant after correction for potential multiple testing errors. In contrast, the *CDKN2B* –593T allele was underrepresented among patients (30.9% versus 37.9%), suggesting a protective effect of this variant (OR = 0.7, 95% CI, 0.6-1.0, $P = .02$). The frequency of carriers of the –593T associated genotypes (AT, TT, and TC) was lower in patients with ALL compared with that in controls, but failed to reach statistical significance (Table 3). In *CDKN1B*, we found that –1608GA heterozygotes were more frequent among patients than

controls (21.4% versus 13.6%), conferring an increased risk of ALL in children (OR = 1.7, 95% CI, 1.0-2.8, $P = .03$), but this result did not remain significant after correction for multiple testing. The initial case-control assessment did not reveal any additional noteworthy associations for the other *CDKN2B* and *CDKN1B* variants, and the frequency of *CDKN1A* variants did not differ between patients and controls (Tables 2-3), indicating that these pSNPs alone do not appear to modify the risk of childhood pre-B ALL in our dataset.

Haplotype analysis

Because single variants in a candidate gene might not be sufficient to capture the genetic variability relative to a given phenotype, promoter haplotypes were constructed for *CDKN2B*, *CDKN1A*, and *CDKN1B*. Haplotype phase was estimated and the corresponding frequencies and distributions among patients with ALL and controls were assessed (Table 5). Omnibus X^2 tests were performed to determine whether the effects of the haplotypes differed significantly between patients and controls. We found evidence for overall association between the 6 *CDKN2B*-derived haplotypes and ALL ($X^2 = 19.1$ [5 degrees of freedom], $P < .001$), which remained significant after multiple testing corrections. When comparing individual haplotype distributions, the largest differences were observed for haplotypes 2B-2, 2B-3, and 2B-6. Haplotype 2B-2 (CTG), carrying the protective -593T allele (Table 2), occurred at a lower frequency among patients with ALL (31.6% versus 37.9%), suggesting an inverse correlation with the disease (OR = 0.8, 95% CI, 0.6-1.0, $P = .04$). In contrast, haplotype 2B-3 (CAG), carrying the high-risk -593A allele, was overrepresented in patients as opposed to in controls (20.2% versus 13.2%) and was significantly associated with an increased risk of ALL (OR = 1.7, 95% CI, 1.2-2.4, $P = .004$). Interestingly, haplotype 2B-6 (TAC) was found exclusively in patients with ALL. For *CDKN1A* and *CDKN1B*, we were able to construct 7 and 5 haplotypes, respectively, but the omnibus tests did not suggest overall

association, nor did any of the individual haplotypes show significant association with the risk of ALL (Table 5).

Family-based analysis

To further assess the impact of these polymorphisms on childhood pre-B ALL risk, we performed a family-based study by genotyping all 10 pSNPs in 148 patient-parental trios. We either analyzed each variant independently (single-marker; Table 6) or as haplotypes (Table 7) using FBATs. In the univariate FBATs, we found a significant preferential transmission of the high-risk *CDKN2A* -222A variant to affected offspring ($Z = 2.60$, $P = .009$; Table 6). The protective *CDKN2B* -593T allele was shown to be transmitted to affected patients less often than expected ($Z = -2.64$, $P = .008$), whereas the high-risk -593A allele was shown to be overtransmitted ($Z = 2.778$, $P = .005$; data not shown). In addition, these findings held true under the dominant and recessive models as well (data not shown). These results, which remained significant after correction for multiple testing, are consistent with the case-control results discussed in “Single-locus analysis” (Tables 2-3). For *CDKN1A* and *CDKN1B*, no single variant was found to be significantly associated with childhood ALL using the additive model (Table 6).

The global haplotype FBATs revealed no significant transmission disequilibrium of the promoter variants of *CDKN2B*, *CDKN1A*, and *CDKN1B* across all trios (Table 7). However FBAT analysis of individual haplotypes showed that *CDKN2B*-1 (CAC), bearing the high-risk A-593 allele, was preferentially transmitted to affected offspring ($Z = 2.22$, $P = .03$), whereas haplotype 2B-2 (CTG) was associated with a protective effect and shown to be undertransmitted to patients with ALL ($Z = -2.44$, $P = .01$). No other significant associations were detected for the *CDKN1A* and *CDKN1B* promoter haplotypes under the additive

model (Table 7). Haplotype CDKN1B-2 (CGG) did show evidence of increased transmission under the recessive model ($Z = 2.44$, $P = .01$), suggesting the possibility that this haplotype carrying the variant -373G is associated with an increased risk of childhood ALL (data not shown). However, it should be noted that though the additive model is expected to perform well even when the true model is nonadditive, misspecification of a recessive or dominant mode of inheritance can lead to a decrease in power of the FBAT (49, 50).

Discussion

CDKIs are key regulators of the G1/S checkpoint (51). Their strict control and concerted action are crucial to maintain cell homeostasis and genomic integrity during cellular division. It is therefore plausible that variation in gene dosage of such critical cell cycle regulators due to functional regulatory polymorphisms could influence cancer susceptibility by altering cell cycle checkpoints. In the present study, we tested this hypothesis in childhood ALL by assessing the genotype and haplotype distributions associated with 10 common pSNPs found in the genes *CDKN2A*, *CDKN2B*, *CDKN1A*, and *CDKN1B*. This genetic epidemiology study was performed in French-Canadians, a population known for its relative genetic homogeneity due to particular demographic and historic characteristics (52). Using 2 distinct yet complementary study designs (case-control and parental trios), we identified putative associations between pSNPs in *CDKN2A* (-222T>A), *CDKN2B* (-593A>T,C), and *CDKN1B* (-1608G>A) and a modified risk of childhood ALL, supporting the idea that DNA variants leading to variable CDKI levels might contribute at least to childhood leukemogenesis.

In the case of *CDKN2A*, earlier work led to the identification of a critical region containing functional promoter activity within the 869 bases immediately

upstream of its coding domain (53). It is conceivable that the variant -222T>A might alter promoter function, perturbing *CDKN2A* expression and therefore cell cycle control. Interestingly, our *in silico* analysis showed that this pSNP leads to the loss of a predicted c-Myb binding site, a transcription factor required for proliferation, differentiation, and survival of hematopoietic cells (54, 55). Furthermore, reduced *CDKN2A* expression due to regulatory polymorphisms has been suggested to contribute to arrested lymphoblast differentiation, which is characteristic of leukemic disorders (56). Additional studies are required however to confirm whether this particular -222T>A nucleotide change has an effect on *CDKN2A* expression. We cannot rule out the possibility of other linked functional SNPs within (or beyond) the *CDKN2A* sequence that could influence childhood ALL predisposition and account for the observed association, especially given the fact that no haplotypic data were available in this particular study. To this effect, variant -222T>A has been shown to be in complete linkage disequilibrium with another common *CDKN2A* variant, an alanine-to-threonine substitution at codon 148 (Ala148Thr) shown to be associated with malignant melanoma (19, 44, 57-59).

For *CDKN2B*, although we failed to show a significant positive association with -593AT heterozygous or -593AA homozygous individuals, the haplotype-specific analysis did support evidence of an association between allele -593A and childhood leukemia. Haplotype 2B-3 (CAG) carrying the -593A allele was overrepresented among patients with ALL, and 2B-1 (CAC) was overtransmitted more frequently from parents to affected offspring, suggesting a potential positive association with the disease. The differential binding detected at the -593A>T,C site in both HeLa and HepG2 nuclear extracts further supports this hypothesis. The fact that these associations were observed for given haplotypes rather than at individual SNPs could reflect the benefit of haplotype-based analysis. Haplotypes provide an advantage because they contain more information as to the genetic variability in the surrounding locus when the

contributing SNPs are not all directly observed, providing they are within linkage units (haplotype blocks) with the SNP under study (60). Using the information generated by the International HapMap Project, we were able to confirm that the *CDKN2B* -593A>T, C variant is found within a 33-kb haplotype block alongside 22 other tagged SNPs. It is possible that at least one of these SNPs in linkage disequilibrium with the -593 variant might contribute to the observed modified risk of childhood ALL.

Both *CDKN2A* and *CDKN2B* are well-characterized tumor suppressors, and their implication in human cancers and various hematologic malignancies, including ALL (61), has been demonstrated. So far, deletion events have been the main cause of somatic inactivation of *CDKN2A/2B* in certain ALL subtypes (62), leading to deregulation of the cell cycle and subsequent tumor genesis. For the first time, we have provided evidence of the implication of *CDKN2A* and *CDKN2B* germline mutations in childhood leukemogenesis.

CDKN1B also plays a critical role in regulating cell proliferation, and studies of knockout mice suggest that *CDKN1B* acts as a tumor suppressor as well (63-66). Furthermore, a familial study on prostate cancer revealed an association with the regulatory SNP rs34330 found in the 5' untranslated region of *CDKN1B*, providing evidence that germline variants of this gene may indeed play a role in cancer susceptibility (18). In this report, we observed an increased risk of childhood ALL among carriers of the *CDKN1B* -1608GA genotype. However, the FBATs failed to corroborate this association, indicating random transmission of the -1608A allele from parents to affected children. To this effect, previous work by Labuda et al. (67) suggested that parental genetics might be important in predicting the risk of cancer (at least childhood leukemia). In other words, if at certain loci the parental genotypes rather than the offspring's own combination of genotypes are responsible for genetic susceptibility to a disease, analyses

such as FBAT would fail to detect the disease-susceptibility allele(s) since the parent-to-child transmission would essentially be random. We tested this hypothesis in our dataset for *CDKN1B* -1608G>A by substituting the parents for the patients and comparing fathers with male controls and mothers to female controls, but we failed to detect any significant associations (data not shown). Although this hypothesis remains speculative and requires further analysis to rule out the possibility that the observed case-control association is simply a spurious result, it illustrates the importance of considering the effect of parental genotypes and of combining various study designs when assessing complex disease susceptibility.

In conclusion, our new findings suggest that germline variants in CDKI genes *CDKN2A*, *CDKN2B*, and *CDKN1B* may play a role in susceptibility to childhood leukemia. Also, we cannot rule out the possibility that these polymorphisms might act in combination with other disease modifiers in the same or any other related biological pathways. Further studies that evaluate interaction effects with genes involved in xenobiotic metabolism and DNA repair will be interesting because they may build upon previous findings of the implication of such disease-susceptibility genes in childhood leukemia and allow us to further understand the biological mechanism underlying the observed associations.

Figures

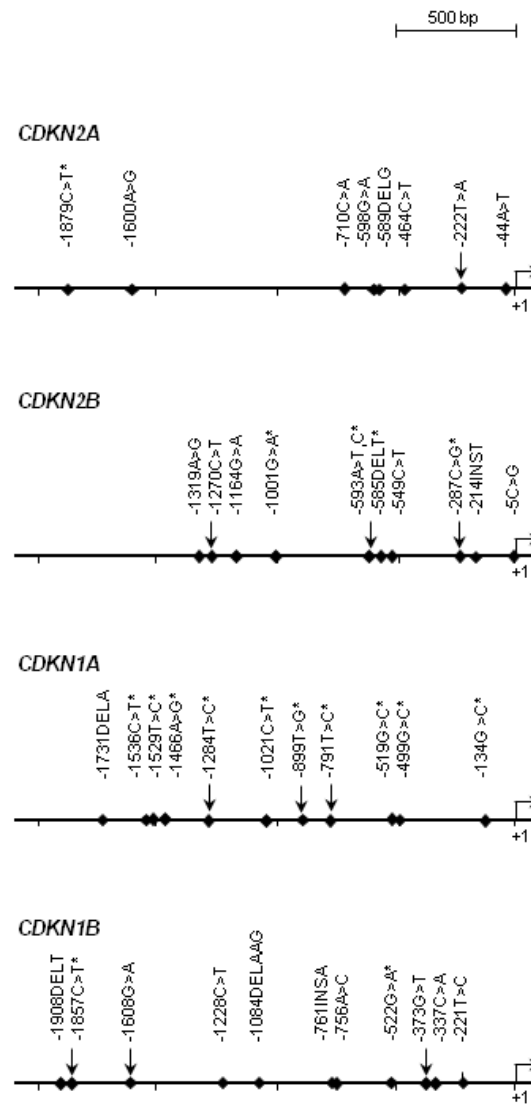


Figure 1. Polymorphisms detected in the promoter regions of *CDKN2A*, *CDKN2B*, *CDKN1A*, and *CDKN1B*

pSNPs that were genotyped in this report are identified by arrows; those reported in public databases are marked by an asterisk. The promoter positions were numbered with respect to the first nucleotide of the first exon as 1, and the nucleotide immediately upstream as -1.

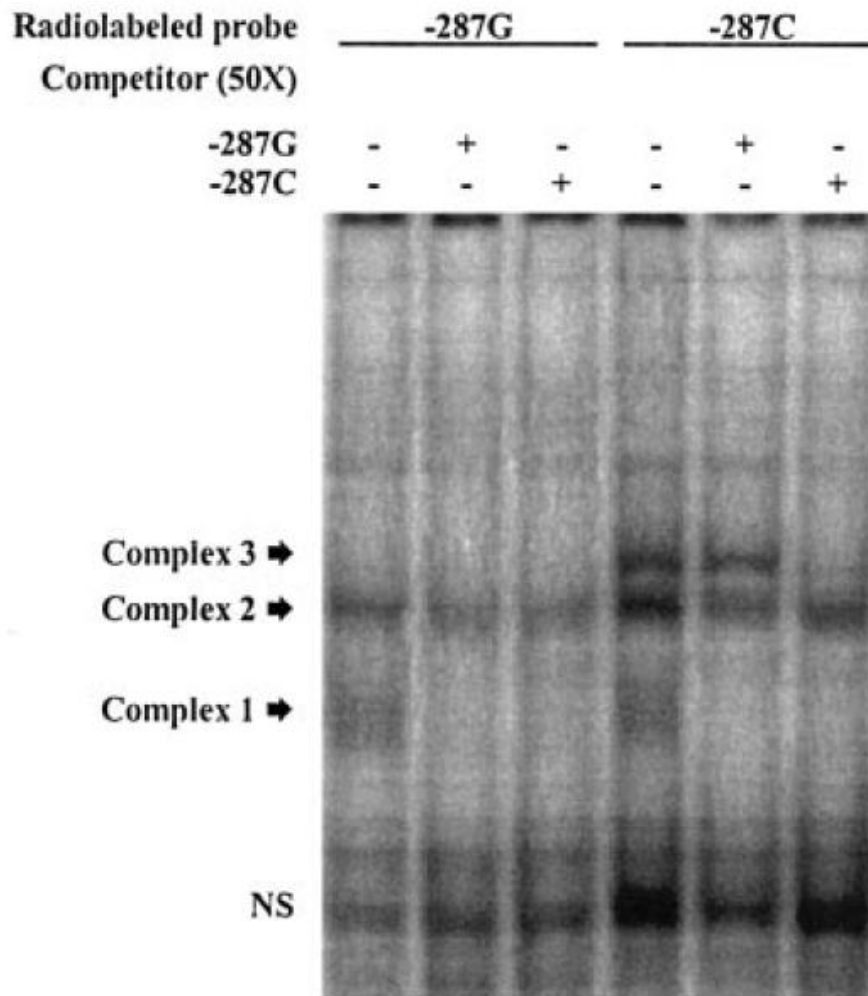


Figure 2. EMSA illustrating allelic DNA-protein interactions in the promoter region of *CDKN2B*

Labeled double-stranded oligonucleotide (ds-oligo) probes corresponding to the *CDKN2B* -287C>G alleles were incubated with HeLa nuclear extracts. Lanes 1 to 3 represent labeled -287G ds-oligos; lanes 4 to 6, labelled -287C ds-oligos. The unlabeled probes used to compete DNA-protein interactions (in 50-fold molar excess) are indicated (+) at the top of each lane. Probe sequences are listed in Table 1. Fast migrating unbound probes can be seen at the bottom of the gel (NS indicates nonspecific), and the position of the DNA-protein complexes of slower mobility are marked by arrows. In this experiment, 3

distinct complexes were found following incubation of the probes with HeLa nuclear extract. Complex 1 was observed with both labeled ds-oligos -287G and -287C (lanes 1 and 4) but was competed by both unlabeled probes, indicating unstable DNA-protein interactions. Complex 2 was also found with both alleles; the -287C-derived complex was competed by both unlabeled probes; however, the -287G-derived complex seemed to be less affected by competitors and thus more stable, suggesting higher binding affinity. Complex 3 appeared only when -287C was present (lane 4 vs lane 1). The specificity of this interaction was illustrated through competition with the specific unlabeled probe, which did not occur with the mismatched -287G probe (lane 6 vs lane 5).

Tables

Table 1. Characteristics of the PCR primers and ASO probes used to genotype the selected pSNPs

Gene (chromosome), DNA variant	SNP ID	PCR primers		Product size, bp	Alleles (ASO probe)	Washing temperature, °C
		Forward	Reverse			
CDKN2A (9p21)						
–222T>A	None	tgttaaaaagaatccgcc	aatgtacacgctgagaaaccc	368	T-222 (ggcagttAggaaggt); A-222 (ggcagttTggaaggt)	45/40*
CDKN2B (9p21)						
–1270C>T	None	gcaacgttttcctaccc	caacagagctgaaatccacc	428	C-1270 (tggacaggGaaaggaac); T-1270 (tggacaggAaaggaac)	55
–593A>T,C	rs2069416	ctgtacaaatgatgaaactggg	ttcactgtggagacgttgg	408	A-593 (gattcttaAagtataat); T-593 (gattcttaTagtataat); C-593 (gattcttaCagtataat)	30
–287G>C	rs2069418	ctgtacaaatgatgaaactggg	ttcactgtggagacgttgg	408	G-287 (ttaagaaaCacggagtt); C-287 (ttaagaaaGacggagtt)	45
CDKN1A (6p21.2)						
–1284T>C	rs733590	ttggatgtataggagcgaagg	aaaaagggcaacctgatctc	897	T-1284 (atcccactAaaaaacag); C-1284 (atcccactGaaaaacag)	45
–899T>G	rs762624	ttggatgtataggagcgaagg	aaaaagggcaacctgatctc	897	T-899 (ggggaaacTggggctca); G-899 (ggggaaacGggggctca)	65
–791T>C	rs2395655	ttggatgtataggagcgaagg	aaaaagggcaacctgatctc	897	T-791 (gaagaaaTccctgtg); C-791 (gaagaaaCccctgtg)	40
CDKN1B (12p13.1–p12)						
–1857C>T	rs3759217	gcattttactggaaaccaacc	tgagacctcttgcttgatcc	663	C-1857 (ctcaagaCagctgca); T-1857 (ctcaagaTagctgca)	40
–1608G>A	None	gcattttactggaaaccaacc	tgagacctcttgcttgatcc	663	G-1680 (tgtccaggGacatgcat); A-1680 (tgtccaggAacatgcat)	52
–373G>T	None	gacaaacaaactagccaaacg	gcagaaaactctgggttaagg	356	G-373 (ttcccgggGaggagc); T-373 (ttcccgggAgaggagc)	60

SNP ID indicates reference SNP identifier number from the NCBI dbSNP database (46). In the ASO probe sequences, uppercase characters indicate the polymorphic site.

*45°C for the T-222 allele, 40°C for the A-222 allele.

Table 2. Allele frequencies of promoter SNPs in *CDKN2A*, *CDKN2B*, *CDKN1A*, and *CDKN1B* in childhood pre-B ALL patients and controls

Genes, DNA variants, and alleles	No. (%)		OR (95% CI)	P
	ALL patients	Controls		
CDKN2A				
–222T>A				
–222T	420 (92.5)	530 (96.4)	1 (referent)	—
–222A	34 (7.5)	20 (3.6)	2.2 (1.2-4.0)	.008*
CDKN2B				
–1270C>T				
–1270C	418 (95.4)	535 (96.9)	1 (referent)	—
–1270T	20 (4.6)	17 (3.1)	1.5 (0.7-3.1)	.24
–593A>T, C				
–593A	289 (66.6)	326 (59.9)	1 (referent)	—
–593T	134 (30.9)	206 (37.9)	0.7 (0.6-1.0)	.02
–593C	11 (2.5)	12 (2.2)	1.0 (0.4-2.6)	.99
–287G>C				
–287G	250 (57.3)	309 (56.4)	1 (referent)	—
–287C	186 (42.7)	239 (43.6)	1.0 (0.7-1.3)	.80
CDKN1A				
–1284T>C				
–1284T	249 (61.9)	352 (64.9)	1 (referent)	—
–1284C	153 (38.1)	190 (35.1)	1.1 (0.9-1.5)	.37
–899T>G				
–899T	298 (75.6)	401 (74.0)	1 (referent)	—
–899G	96 (24.4)	141 (26.0)	0.9 (0.7-1.3)	.59
–791T>C				
–791T	237 (59.0)	331 (61.1)	1 (referent)	—
–791C	165 (41.0)	211 (38.9)	1.1 (0.9-1.5)	.46
CDKN1B				
–1857C>T				
–1857C	371 (90.1)	470 (89.4)	1 (referent)	—
–1857T	41 (10.0)	56 (10.7)	0.9 (0.6-1.5)	.75
–1608G>A				
–1608G	418 (89.3)	483 (91.8)	1 (referent)	—
–1608A	50 (10.7)	43 (8.2)	1.3 (0.9-2.1)	.19
–373G>T				
–373G	261 (59.3)	316 (57.0)	1 (referent)	—
–373T	179 (40.7)	238 (43.0)	0.9 (0.7-1.2)	.48

Percentages indicate number of chromosomes with given allele/total number of chromosomes. OR indicates crude odds ratio; —, not applicable.

*Remained significant following multiple test correction with an FDR of 10%.

Table 3. Distribution of *CDKN2A*, *CDKN2B*, *CDKN1A*, and *CDKN1B* promoter-based genotypes among childhood patients with pre-B ALL and controls

Genes, DNA variant, and genotype	No. (%)		OR (95% CI)	P
	ALL patients	Controls		
CDKN2A				
–222T>A				
TT	195 (85.9)	256 (93.1)	1 (referent)	—
TA	30 (13.2)	18 (6.5)	2.6 (1.1-4.3)	.01
AA	2 (0.9)	1 (0.4)	2.6 (0.1-155.5)	.58
CDKN2B				
–1270C>T				
CC	200 (91.3)	260 (94.2)	1 (referent)	—
CT	18 (8.2)	15 (5.4)	1.6 (0.7-3.4)	.28
TT	1 (0.5)	1 (0.4)	1.3 (0.02-102.4)	.99
–593A>T,C				
AA	97 (44.7)	98 (36.0)	1 (referent)	—
AT	86 (39.6)	124 (45.6)	0.7 (0.5-1.1)	.09
TT	23 (10.6)	38 (14.0)	0.6 (0.3-1.1)	.11
AC	9 (4.2)	6 (2.2)	1.5 (0.5-5.4)	.59
CC	0 (0)	0 (0)	—	—
TC	2 (0.9)	6 (2.2)	0.3 (0.03-2.0)	.28
–287G>C				
GG	73 (33.5)	82 (30.0)	1 (referent)	—
GC	104 (47.7)	145 (52.9)	0.8 (0.5-1.2)	.30
CC	41 (18.8)	47 (17.1)	1.0 (0.6-1.7)	.99
CDKN1A				
–1284T>C				
TT	76 (37.8)	110 (40.6)	1 (referent)	—
TC	97 (48.3)	132 (48.7)	1.1 (0.7-1.6)	.77
CC	28 (13.9)	29 (10.7)	1.4 (0.7-2.7)	.29
–899T>G				
TT	114 (57.9)	149 (55.0)	1 (referent)	—
TG	70 (35.5)	103 (38.0)	0.9 (0.6-1.3)	.62
GG	13 (6.6)	19 (7.0)	0.9 (0.4-2.0)	.85
–791T>C				
TT	70 (34.8)	99 (36.5)	1 (referent)	—
TC	97 (48.3)	133 (49.1)	1.0 (0.7-1.6)	.92
CC	34 (16.9)	39 (14.4)	1.2 (0.7-2.2)	.48
CDKN1B				
–1857C>T				
CC	166 (80.6)	209 (79.5)	1 (referent)	—
CT	39 (18.9)	52 (19.8)	0.9 (0.6-1.5)	.91
TT	1 (0.5)	2 (0.7)	0.6 (0.01-12.2)	.99
–1608G>A				
GG	184 (78.6)	231 (84.9)	1 (referent)	—
GA	50 (21.4)	37 (13.6)	1.7 (1.0-2.8)	.03
AA	0 (0)	4 (1.5)	—	—
–373G>T				
GG	77 (35.0)	89 (32.1)	1 (referent)	—
GT	107 (48.6)	138 (49.8)	0.9 (0.6-1.4)	.61
TT	36 (16.4)	50 (18.1)	0.8 (0.5-1.5)	.51

Percentages indicate number of individuals with a given genotype/total number of genotyped individuals. OR indicates crude odds ratio; —, not applicable.

Table 4. Impact of promoter SNPs on predicted transcription factor binding sites

Gene, pSNP	In silico predictions for TF binding sites		Validated differential binding		
	Gains (score*)	Losses (score)	HeLa	Jeg-3	HepG2
CDKN2A					
–222T>A	—	c-Myb (> 0.97), myogenin/nuclear factor 1 (> 0.74)	No	NA	No
CDKN2B					
–1270C>T	Interferon regulatory factor 2 (> 0.79), Pax1 paired domain protein (> 0.64)	B-cell-specific activating protein (> 0.79)	No	Yes	NA
–593A>T,C†	—	MyT1 zinc finger transcription factor (> 0.80), prostate-specific homeodomain protein NKX3.1 (0.84)	Yes	No	Yes
–287G>C	—	Neural-restrictive silencer element (0.70)	Yes	No	No
CDKN1A					
–1284T>C	—	—	No	Yes	Yes
–899T>G	v-Myb (> 0.92)	—	Yes	Yes	No
–791T>C	—	—	No	Yes	No
CDKN1B					
–1857C>T	Ecotropic viral integration site 1 encoded factor (> 0.83)	Myoblast determining factor (> 0.83)	No	Yes	No
–1608G>A	POZ/zinc finger protein (> 0.84)	—	No	No	No
–373G>T	GAGA-Box (> 0.78), olfactory neuron-specific factor (> 0.82), Pax1 paired domain protein (> 0.63), signal transducer and activator of transcription 1 (> 0.81), STAT6 (> 0.84)	Myc-associated zinc finger protein (> 0.94)	No	No	No

Predicted impact of the pSNPs was estimated with the program matInspector. Gain and loss of transcription factor binding sites are specified for the minor allele with respect to the major allele. NA indicates not available; —, none.

*Matrix similarity score.

†Predictions for the minor allele –593T only are shown. *In silico* analysis for allele –593C was inconclusive.

Table 5. Distribution of *CDKN2B*, *CDKN1A*, and *CDKN1B* promoter haplotypes in patients with pre-B ALL and controls

Gene, Haplotype	DNA Variant			No. (%)		OR (95% CI)	<i>P</i>
	Variant 1	Variant 2	Variant 3	ALL patients	Controls		
<i>CDKN2B</i>*	–1270C>T	–593A>T,C	–287G>C				
2B-1	C	A	C	189 (41.4)	242 (43.7)	0.9 (0.7-1.2)	.48
2B-2	C	T	G	144 (31.6)	210 (37.9)	0.8 (0.6-1.0)	.04
2B-3	C	A	G	92 (20.2)	73 (13.2)	1.7 (1.2-2.4)	.004§
2B-4	T	A	G	13 (2.9)	17 (3.1)	0.9 (0.4-2.0)	> .999
2B-5	C	C	G	11 (2.4)	12 (2.2)	1.1 (0.4-2.8)	.83
2B-6	T	A	C	7 (1.5)	0 (0)	—	—
<i>CDKN1A</i>†	–1284T>C	–899T>G	–791T>C				
1A-1	T	T	T	234 (57.9)	322 (59.4)	1.0 (0.8-1.3)	> .999
1A-2	C	G	C	78 (19.3)	111 (20.5)	1.0 (0.7-1.3)	.81
1A-3	C	T	C	71 (17.6)	76 (14.0)	1.3 (0.9-1.9)	.10
1A-4	T	G	C	16 (4.0)	24 (4.4)	0.9 (0.4-1.8)	.87
1A-5	T	G	T	1 (0.2)	6 (1.1)	0.2 (0.004-1.9)	.25
1A-6	C	T	T	3 (0.7)	3 (0.5)	1.4 (0.2-10.3)	.70
1A-7	C	G	T	1 (0.2)	0 (0)	—	—
<i>CDKN1B</i>‡	–1857C>T	–1608G>A	–373G>T				
1B-1	C	G	T	189 (40.4)	238 (43.0)	0.9 (0.7-1.1)	.25
1B-2	C	G	G	190 (40.4)	215 (38.8)	1.0 (0.8-1.3)	.85
1B-3	C	A	G	50 (10.6)	45 (8.1)	1.3 (0.8-2.1)	.23
1B-4	T	G	G	40 (8.5)	56 (10.1)	0.8 (0.5-1.3)	.33
1B-5	T	G	T	1 (0.2)	0 (0)	—	—

The risk of ALL was evaluated for each haplotype compared with all other possible haplotypes combined. The X^2 values represent overall haplotype frequency comparisons between ALL patients and control subjects for each gene.

Percentages indicate number of chromosomes with given haplotype/total number of chromosomes. OR indicates crude odds ratio; df, degrees of freedom; and —, not applicable.

$$*X_{df=5}^2 = 19.1, P < .001.$$

$$\dagger X_{df=6}^2 = 5.0, P = .59.$$

$$\ddagger X_{df=4}^2 = 3.7, P = .44.$$

§Remained significant following multiple test correction with an FDR of 10%.

Table 6. Family-based association analysis of promoter SNPs in *CDKN2A*, *CDKN2B*, *CDKN1A*, and *CDKN1B*

Gene and SNP	Variant	MAF	No. of families*	Statistic†	E(S)‡	Var(S)§	Z	P
<i>CDKN2A</i>								
-222T>A	A	.06	25	19	12.5	6.3	2.60	.009¶
<i>CDKN2B</i>								
-1270C>T	T	.05	16	6	8.0	4.0	-1.00	.32
-593A>T,C	T/C	.33/.02	81/11	53/5	66.5/6.0	26.3/3.0	-2.64/-0.58	.008¶/0.56
-287G>C	C	.45	92	94	85.5	30.3	1.55	.12
<i>CDKN1A</i>								
-1284T>C	C	.39	72	66	64.5	24.3	0.31	.76
-899T>G	G	.29	62	45	46.5	19.8	-0.34	.74
-791T>C	C	.43	72	70	69.5	24.8	0.10	.92
<i>CDKN1B</i>								
-1857C>T	T	.11	39	17	21.5	10.8	-1.37	.17
-1608G>A	A	.10	41	26	22.5	10.8	1.07	.29
-373G>T	T	.41	84	66	73.5	29.3	-1.39	.17

FBAT analyses were performed under the additive model. Only the results for the minor alleles are shown. MAF indicates minor (i.e., variant) allele frequency.

*Number of informative families (i.e., families with at least one heterozygote parent).

†Test statistic from FBAT for the observed number of transmitted alleles.

‡Expected value of S under the null hypothesis (i.e., no linkage or association).

§Variance of the test statistic S.

¶Remained significant following multiple test correction with an FDR of 10%.

Table 7. FBAT analysis of promoter haplotypes in *CDKN2B*, *CDKN1A*, and *CDKN1B*

Gene and haplotype	Frequency	No. of families*	Statistic†	E(S)‡	Var(S)§	Z	P	χ^2 (df)	Global P
<i>CDKN2B</i>								8.0 (4)	.09
CAC	.41	77.0	88.1	76.5	27.1	2.22	.03		
CTG	.35	69.9	56.9	69.9	28.5	-2.44	.01		
CAG	.18	61.0	44.9	40.5	16.6	1.09	.27		
TAG	.03	10.2	3.2	5.1	2.4	-1.24	.22		
CCG	.02	9.0	—	—	—	—	—		
TAC	.01	4.8	—	—	—	—	—		
<i>CDKN1A</i>								3.9 (4)	.42
TTT	.55	59.0	75	76.5	24.3	-0.31	.76		
CGC	.22	47.0	35.2	38.1	16.9	-0.71	.48		
CTC	.17	45.0	36.8	31.9	13.5	1.33	.18		
TGC	.05	19.8	11.8	10.4	5.7	0.59	.56		
TGT	.01	2.0	—	—	—	—	—		
CTT	.01	3.0	—	—	—	—	—		
CGT	.002	1.0	—	—	—	—	—		
<i>CDKN1B</i>								5.9 (4)	.21
CGT	.42	75.9	74.8	80.7	27.0	-1.15	.25		
CGG	.38	76.0	80.7	73.5	24.1	1.47	.14		
CAG	.09	35.2	22.4	19.1	8.4	1.14	.25		
TGG	.11	35.8	16.4	21.2	10.2	-1.49	.14		
TGT	.01	2.3	—	—	—	—	—		

Haplotype-specific FBAT analyses were performed under the additive model. df indicates degrees of freedom; —, not applicable due to lack of informative families (i.e., < 10).

*Number of informative families (i.e., families with at least 1 heterozygote parent).

†Test statistic from FBAT for the observed number of transmitted alleles.

‡Expected value of S under the null hypothesis (i.e., no linkage or association).

§Variance of the test statistic S.

References

1. Stiller CA. Epidemiology and genetics of childhood cancer. *Oncogene*. 2004 Aug 23;23(38):6429-44.
2. Alves S, Amorim A, Ferreira F, Norton L, Prata MJ. The GSTM1 and GSTT1 genetic polymorphisms and susceptibility to acute lymphoblastic leukemia in children from north Portugal. *Leukemia*. 2002 Aug;16(8):1565-7.
3. Canalle R, Burim RV, Tone LG, Takahashi CS. Genetic polymorphisms and susceptibility to childhood acute lymphoblastic leukemia. *Environ Mol Mutagen*. 2004;43(2):100-9.
4. Chen CL, Liu Q, Pui CH, Rivera GK, Sandlund JT, Ribeiro R, et al. Higher frequency of glutathione S-transferase deletions in black children with acute lymphoblastic leukemia. *Blood*. 1997;89(5):1701-7.
5. Joseph T, Kusumakumary P, Chacko P, Abraham A, Radhakrishna Pillai M. Genetic polymorphism of CYP1A1, CYP2D6, GSTM1 and GSTT1 and susceptibility to acute lymphoblastic leukaemia in Indian children. *Pediatr Blood Cancer*. 2004 Oct;43(5):560-7.
6. Krajcinovic M, Labuda D, Richer C, Karimi S, Sinnett D. Susceptibility to childhood acute lymphoblastic leukemia: influence of CYP1A1, CYP2D6, GSTM1, and GSTT1 genetic polymorphisms. *Blood*. 1999 1999;93(5):1496-501.
7. Krajcinovic M, Labuda D, Sinnett D. Glutathione S-transferase P1 genetic polymorphisms and susceptibility to childhood acute lymphoblastic leukaemia. *Pharmacogenetics*. 2002 Nov;12(8):655-8.
8. Krajcinovic M, Richer C, Sinnett H, Labuda D, Sinnett D. Genetic polymorphisms of N-acetyltransferases 1 and 2 and gene-gene interaction in the susceptibility to childhood acute lymphoblastic leukemia. *Cancer Epidemiol Biomarkers Prev*. 2000 2000;9(6):557-62.
9. Krajcinovic M, Sinnett H, Richer C, Labuda D, Sinnett D. Role of NQO1, MPO and CYP2E1 genetic polymorphisms in the susceptibility to childhood acute lymphoblastic leukemia. *Int J Cancer*. 2002 2002;97(2):230-6.
10. Pakakasama S, Mukda E, Sasanakul W, Kadegasem P, Udomsubpayakul U, Thithapandha A, et al. Polymorphisms of drug-metabolizing enzymes and risk of childhood acute lymphoblastic leukemia. *Am J Hematol*. 2005 Jul;79(3):202-5.
11. Wiemels JL, Smith RN, Taylor GM, Eden OB, Alexander FE, Greaves MF. Methylenetetrahydrofolate reductase (MTHFR) polymorphisms and risk of

molecularly defined subtypes of childhood acute leukemia. *Proc Natl Acad Sci U S A*. 2001;98(7):4004-9.

12. Smith MT, Wang Y, Skibola CF, Slater DJ, Lo Nigro L, Nowell PC, et al. Low NAD(P)H:quinone oxidoreductase activity is associated with increased risk of leukemia with MLL translocations in infants and children. *Blood*. 2002 Dec 15;100(13):4590-3.

13. Joseph T, Kusumakumary P, Chacko P, Abraham A, Pillai MR. DNA repair gene XRCC1 polymorphisms in childhood acute lymphoblastic leukemia. *Cancer Lett*. 2005 Jan 10;217(1):17-24.

14. Mathonnet G, Krajcinovic M, Labuda D, Sinnott D. Role of DNA mismatch repair genetic polymorphisms in the risk of childhood acute lymphoblastic leukaemia. *Br J Haematol*. 2003 Oct;123(1):45-8.

15. Malumbres M, Barbacid M. To cycle or not to cycle: a critical decision in cancer. *Nat Rev Cancer*. 2001 Dec;1(3):222-31.

16. Ivanchuk SM, Rutka JT. The cell cycle: accelerators, brakes, and checkpoints. *Neurosurgery*. 2004 Mar;54(3):692-9; discussion 9-700.

17. Johnson DG, Walker CL. Cyclins and cell cycle checkpoints. *Annu Rev Pharmacol Toxicol*. 1999;39:295-312.

18. Chang BL, Zheng SL, Isaacs SD, Wiley KE, Turner A, Li G, et al. A polymorphism in the CDKN1B gene is associated with increased risk of hereditary prostate cancer. *Cancer Res*. 2004 Mar 15;64(6):1997-9.

19. Debniak T, Scott RJ, Huzarski T, Byrski T, Rozmiarek A, Debniak B, et al. CDKN2A common variants and their association with melanoma risk: a population-based study. *Cancer Res*. 2005 Feb 1;65(3):835-9.

20. Harland M, Taylor CF, Bass S, Churchman M, Randerson-Moor JA, Holland EA, et al. Intronic sequence variants of the CDKN2A gene in melanoma pedigrees. *Genes Chromosomes Cancer*. 2005 Jun;43(2):128-36.

21. Kibel AS, Suarez BK, Belani J, Oh J, Webster R, Brophy-Ebbers M, et al. CDKN1A and CDKN1B polymorphisms and risk of advanced prostate carcinoma. *Cancer Res*. 2003 May 1;63(9):2033-6.

22. Rodrigues FC, Kawasaki-Oyama RS, Fo JF, Ukuyama EE, Antonio JR, Bozola AR, et al. Analysis of CDKN1A polymorphisms: markers of cancer susceptibility? *Cancer Genet Cytogenet*. 2003 Apr 15;142(2):92-8.

23. Sauroja I, Smeds J, Vlaykova T, Kumar R, Talve L, Hahka-Kemppinen M, et al. Analysis of G(1)/S checkpoint regulators in metastatic melanoma. *Genes Chromosomes Cancer*. 2000 Aug;28(4):404-14.

24. Lukas J, Groshen S, Saffari B, Niu N, Reles A, Wen WH, et al. WAF1/Cip1 gene polymorphism and expression in carcinomas of the breast, ovary, and endometrium. *Am J Pathol*. 1997 Jan;150(1):167-75.
25. Banerjee P, Bahlo M, Schwartz JR, Loots GG, Houston KA, Dubchak I, et al. SNPs in putative regulatory regions identified by human mouse comparative sequencing and transcription factor binding site data. *Mamm Genome*. 2002 Oct;13(10):554-7.
26. Pastinen T, Hudson TJ. Cis-acting regulatory variation in the human genome. *Science*. 2004 Oct 22;306(5696):647-50.
27. Yasuda T, Takeshita H, Nakazato E, Nakajima T, Nakashima Y, Mori S, et al. The molecular basis for genetic polymorphism of human deoxyribonuclease II (DNase II): a single nucleotide substitution in the promoter region of human DNase II changes the promoter activity. *FEBS Lett*. 2000 Feb 11;467(2-3):231-4.
28. Li LC, Chui RM, Sasaki M, Nakajima K, Perinchery G, Au HC, et al. A single nucleotide polymorphism in the E-cadherin gene promoter alters transcriptional activities. *Cancer Res*. 2000 Feb 15;60(4):873-6.
29. Smeraldi E, Zanardi R, Benedetti F, Di Bella D, Perez J, Catalano M. Polymorphism within the promoter of the serotonin transporter gene and antidepressant efficacy of fluvoxamine. *Mol Psychiatry*. 1998 Nov;3(6):508-11.
30. Sanak M, Pierzchalska M, Bazan-Socha S, Szczeklik A. Enhanced expression of the leukotriene C(4) synthase due to overactive transcription of an allelic variant associated with aspirin-intolerant asthma. *Am J Respir Cell Mol Biol*. 2000 Sep;23(3):290-6.
31. Pihlajamäki J, Karjalainen L, Karhapää P, Vauhkonen I, Taskinen MR, Deeb SS, et al. G-250A substitution in promoter of hepatic lipase gene is associated with dyslipidemia and insulin resistance in healthy control subjects and in members of families with familial combined hyperlipidemia. *Arterioscler Thromb Vasc Biol*. 2000 Jul;20(7):1789-95.
32. Lopez ER, Zwermann O, Segni M, Meyer G, Reinke M, Seissler J, et al. A promoter polymorphism of the CYP27B1 gene is associated with Addison's disease, Hashimoto's thyroiditis, Grave's disease and type 1 diabetes mellitus in Germans. *Eur J Endocrinol*. 2004;151(2):193-7.
33. Ho SY, Wang YJ, Chen HL, Chen CH, Chang CJ, Wang PJ, et al. Increased risk of developing hepatocellular carcinoma associated with carriage of the TNF2 allele of the -308 tumor necrosis factor- α promoter gene. *Cancer Causes Control*. 2004 September;15(7):657-63.
34. Baccichet A, Qualman SK, Sinnett D. Allelic loss in childhood acute lymphoblastic leukemia. *Leuk Res*. 1997 1997;21(9):817-23.

35. Bourgeois S, Labuda D. Dynamic allele-specific oligonucleotide hybridization on solid support (DASO). *Anal Biochem.* 2004;324:309-11.
36. Labuda D, Krajcinovic M, Richer C, Skoll A, Sinnett H, Yotova V, et al. Rapid detection of CYP1A1, CYP2D6, and NAT variants by multiplex polymerase chain reaction and allele-specific oligonucleotide assay. *Anal Biochem.* 1999 1999;275(1):84-92.
37. Belanger H, Beaulieu P, Moreau C, Labuda D, Hudson TJ, Sinnett D. Functional promoter SNPs in cell cycle checkpoint genes. *Hum Mol Genet.* 2005 Sep 15;14(18):2641-8.
38. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 2001 2001;68(4):978-89.
39. Excoffier L, Laval G, Schneider S. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online.* 2005;1:47-50.
40. Seltman H, Roeder K, Devlin B. Evolutionary-based association analysis using haplotype data. *Genet Epidemiol.* 2003 Jul;25(1):48-58.
41. Murphy JM, Laird NM, Monson RR, Sobol AM, Leighton AH. Incidence of depression in the Stirling County Study: historical and comparative perspectives. *Psychol Med.* 2000 May;30(3):505-14.
42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical approach and powerful approach for multiple testing. *J R Stat Soc B.* 1995;57:289-300.
43. Sinnett D, Beaulieu P, Belanger H, Lefebvre JF, Langlois S, Theberge MC, et al. Detection and characterization of DNA variants in the promoter regions of hundreds of human disease candidate genes. *Genomics.* 2006 Jun;87(6):704-10.
44. Harland M, Holland EA, Ghiorzo P, Mantelli M, Bianchi-Scarra G, Goldstein AM, et al. Mutation screening of the CDKN2A promoter in melanoma families. *Genes Chromosomes Cancer.* 2000 May;28(1):45-57.
45. Pollock PM, Stark MS, Palmer JM, Walters MK, Aitken JF, Martin NG, et al. Mutation analysis of the CDKN2A promoter in Australian melanoma families. *Genes Chromosomes Cancer.* 2001 Sep;32(1):89-94.
46. National Center for Biotechnology Information. dbSNP home page. <http://www.ncbi.nlm.nih.gov/projects/SNP/>. Accessed September 2005.
47. International HapMap project. <http://www.hapmap.org>. Accessed September 2005.

48. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, et al. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*. 2005 Jul 1;21(13):2933-42.
49. Knapp M. A note on power approximations for the transmission/disequilibrium test. *Am J Hum Genet*. 1999 Apr;64(4):1177-85.
50. Tu IP, Balise RR, Whittemore AS. Detection of disease genes by use of family data. II. Application to nuclear families. *Am J Hum Genet*. 2000 Apr;66(4):1341-50.
51. Sherr CJ, Roberts JM. CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes Dev*. 1999 Jun 15;13(12):1501-12.
52. Labuda D, Zietkiewicz E, Labuda M. The genetic clock and the age of the founder effect in growing populations: a lesson from French Canadians and Ashkenazim. *Am J Hum Genet*. 1997 1997;61(3):768-71.
53. Hara E, Smith R, Parry D, Tahara H, Stone S, Peters G. Regulation of p16CDKN2 expression and its implications for cell immortalization and senescence. *Mol Cell Biol*. 1996 Mar;16(3):859-67.
54. Ezoe S, Matsumura I, Satoh Y, Tanaka H, Kanakura Y. Cell cycle regulation in hematopoietic stem/progenitor cells. *Cell Cycle*. 2004 Mar;3(3):314-8.
55. Golay J, Basilico L, Loffarelli L, Songia S, Broccoli V, Introna M. Regulation of hematopoietic cell proliferation and differentiation by the myb oncogene family of transcription factors. *Int J Clin Lab Res*. 1996;26(1):24-32.
56. Furukawa Y, Kikuchi J, Nakamura M, Iwase S, Yamada H, Matsuda M. Lineage-specific regulation of cell cycle control gene expression during haematopoietic cell differentiation. *Br J Haematol*. 2000 Sep;110(3):663-73.
57. Debniak T, Gorski B, Scott RJ, Cybulski C, Medrek K, Zowocka E, et al. Germline mutation and large deletion analysis of the CDKN2A and ARF genes in families with multiple melanoma or an aggregation of malignant melanoma and breast cancer. *Int J Cancer*. 2004 Jul 1;110(4):558-62.
58. Lamperska K, Karezewska A, Kwiatkowska E, Mackiewicz A. Analysis of mutations in the p16/CDKN2A gene in sporadic and familial melanoma in the Polish population. *Acta Biochim Pol*. 2002;49(2):369-76.
59. Mantelli M, Barile M, Ciotti P, Ghiorzo P, Lantieri F, Pastorino L, et al. High prevalence of the G101W germline mutation in the CDKN2A (P16(ink4a)) gene in 62 Italian malignant melanoma families. *Am J Med Genet*. 2002 Jan 22;107(3):214-21.
60. Judson R, Stephens JC. Notes from the SNP vs. haplotype front. *Pharmacogenomics*. 2001 2001;2(1):7-10.

61. Drexler HG. Review of alterations of the cyclin-dependent kinase inhibitor INK4 family genes p15, p16, p18 and p19 in human leukemia-lymphoma cells. *Leukemia*. 1998 Jun;12(6):845-59.
62. Bertin R, Acquaviva C, Mirebeau D, Guidal-Giroux C, Vilmer E, Cave H. CDKN2A, CDKN2B, and MTAP gene dosage permits precise characterization of mono- and bi-allelic 9p21 deletions in childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer*. 2003 May;37(1):44-57.
63. Kiyokawa H, Kineman RD, Manova-Todorova KO, Soares VC, Hoffman ES, Ono M, et al. Enhanced growth of mice lacking the cyclin-dependent kinase inhibitor function of p27(Kip1). *Cell*. 1996 May 31;85(5):721-32.
64. Nakayama K, Ishida N, Shirane M, Inomata A, Inoue T, Shishido N, et al. Mice lacking p27(Kip1) display increased body size, multiple organ hyperplasia, retinal dysplasia, and pituitary tumors. *Cell*. 1996 May 31;85(5):707-20.
65. Fero ML, Randel E, Gurley KE, Roberts JM, Kemp CJ. The murine gene p27Kip1 is haplo-insufficient for tumour suppression. *Nature*. 1998 Nov 12;396(6707):177-80.
66. Shaffer DR, Viale A, Ishiwata R, Leversha M, Olgac S, Manova K, et al. Evidence for a p27 tumor suppressive function independent of its role regulating cell proliferation in the prostate. *Proc Natl Acad Sci U S A*. 2005 Jan 4;102(1):210-5.
67. Labuda D, Krajcinovic M, Sabbagh A, Infante-Rivard C, Sinnett D. Parental genotypes in the risk of a complex disease. *Am J Hum Genet*. 2002 2002;71(1):193-7.

Reprint Permissions



Institution: UNIVERSITE DE MONTREAL | [Sign In via User Name/Password](#)

SEARCH:

GO

Advanced

Current Issue

First Edition

e-Letters

Future Articles

Archives

Submit to Blood

Search Blood

ASH™

Meeting Abstracts

E-Mail Alerts

Rights & Permissions

Material published in *Blood* is covered by copyright. All rights reserved. No part of the publication may be reproduced (see exception below), stored in a retrieval system, translated, or transmitted in any form or by any means now or hereafter known, electronic or mechanical, without permission in writing from the publisher, The American Society of Hematology (ASH).

The copyright owner consents that copies of articles may be made for personal or internal use, or for the personal or internal use of specific clients, for those registered with the [Copyright Clearance Center Inc.](#) This consent is given on the condition that the copier pay the stated per-copy fees through the Copyright Clearance Center, Inc., for copying beyond that permitted by Sections 107 and 108 of the US Copyright Law (Fair Use). This consent does not extend to other kinds of copying, such as copying for general distribution, for advertising or for promotional purposes, for creating new collective works, or for resale. For those kinds of purposes, permission must be sought from the publisher using the following guidelines. If you have any questions regarding our rights and permissions policies, or if you have a special request not covered by these guidelines, please contact [Blood Permissions](#).

Republication Requests

- [Authors reusing their own material](#)

- [All others wishing to use material](#)

[Photocopy Requests](#)

[Translation Rights](#)

[Regulatory Approval/Patent and Trademark Applications](#)

[Reprints](#)

- [Author reprints](#)

- [Commercial reprints](#)

[ASH Annual Meeting abstracts](#)

[Blood Embargo Policy](#)

[Copyright Clearance Center](#)

Republication Requests

Authors reusing their own material

Authors have permission to do the following after their article has been published in *Blood*, either in print or online as a First Edition Paper.

- Reprint the article in print collections of the author's own writing.
- Present the work orally in its entirety.
- Use the article in theses and/or dissertation.
- Reproduce the article for use in courses the author is teaching. If the author is employed by an academic institution, that institution may also reproduce the article for course teaching.
- Distribute photocopies of the article to colleagues, but only for noncommercial purposes.
- Reuse figures and tables created by the author in future works.
- Post a copy of the article on the author's personal website, departmental website, and/or the university intranet. A hyperlink to the article on the *Blood* website must be included.

The author must include the following citation when citing material that appeared in the print edition of *Blood*:

"This research was originally published in *Blood*. Author(s). Title. *Blood*. Year;Vol:pp-pp. © the American Society of Hematology."



No Evidence for Association Between *TGFB1* Promoter SNPs and the Risk of Childhood Pre-B Acute Lymphoblastic Leukemia Among French Canadians

Jasmine Healy,¹ Marie-Helene Roy-Gagnon,^{1,2} and Daniel Sinnett^{1,3}

¹Division of Hematology-Oncology, Sainte-Justine Hospital Research Center, University of Montreal, QC, Canada;

²Department of Social and Preventive Medicine, Faculty of Medicine, University of Montreal, QC, Canada;

³Department of Pediatrics, Faculty of Medicine, University of Montreal, QC, Canada

This research was originally published in *Haematologica the Hematology Journal*. Healy J, Roy-Gagnon MH, Sinnett D. No evidence for association between *TGFB1* promoter SNPs and the risk of childhood pre-B acute lymphoblastic leukemia among French Canadians. *Haematologica*. 2009 Jul;94(7):1034-5. © the Ferrata Storti Foundation. The article is included with the permission of the authors and the journal.

Author Contributions

For this article, I carried out 90% of the work including the genotyping and statistical analyses. I also wrote the manuscript. MH Roy-Gagnon participate in the analysis and design of the study. D. Sinnett is principle investigator and participated in the design and coordination of the study as well as the writing of the manuscript. All authors read and approved the final manuscript.

Letter to the Editor

TGFB1 is a potent growth inhibitor to a wide variety of cell types including hematopoietic cells and deregulation of the TGFB1 signaling pathway has been implicated in the development of several cancers (1). Functional regulatory polymorphisms of the *TGFB1* gene have been directly associated with interindividual variability in TGFB1 plasma levels and modified risk of breast (2), lung (3), colorectal (4), and prostate (5) cancers. No studies, to date, have examined the association between genetic polymorphisms in *TGFB1* and childhood leukemia. In this study we examined the role of *TGFB1* promoter SNPs (pSNPs), -1886G>A, -1571A>G, -1550DEL/AGG, and -509C>T, as genetic modulators of childhood pre-B ALL susceptibility among the French Canadian population.

In a case-control study, we investigated genotypic, haplotypic, as well as multi-SNP combination associations with childhood pre-B ALL. The study population and inclusion criteria were described previously (6). Childhood pre-B ALL cases (n=321) consisted of 189 boys and 132 girls with a median age of 4.7 years. Parental DNA was available for 203 of these children. Healthy controls (n=329) were recruited at the Sainte-Justine Hospital. Study individuals were all French Canadian from the province of Quebec, Canada. Candidate pSNPs were previously identified (7) and were selected based on their frequency in the European population. A PCR-based allele-specific oligonucleotide hybridization approach was used to genotype samples, described previously (8).

Hardy-Weinberg equilibrium was tested using the X^2 goodness of fit test and PedCheck (Version 1.1) was used to identify genotype incompatibilities using

the familial data (9). Pearson's X^2 test or Fisher's exact test, as appropriate, was used to compare allele/genotype/haplotype carriership in patients and controls. Crude odds ratios (ORs) were measured using logistic regression in STATA (Release 9.2) and are given with 95% confidence intervals (CIs). Haplotype frequencies were estimated using the FAMHAP Software (Version 16) using parental data when available (10). A likelihood ratio test implemented in FAMHAP was used to examine global haplotype associations with disease status. Multimarker combinations were tested for association with disease using the method proposed by Becker and Knapp implemented in FAMHAP (11). Multiple testing was controlled for using the false discovery rate (12) with a type I error rate of 10%.

Frequencies of pSNPs were in agreement with those previously reported in other populations of European descent and all distributions were in Hardy-Weinberg equilibrium. *TGFB1* pSNP distributions did not differ significantly between cases and controls (Table 1). Suggestive evidence of a reduced risk of childhood pre-B ALL was demonstrated when carriers of at least one variant -1886A allele were compared to homozygous individuals for the ancestral -1886G allele (OR=0.58, 95% CI: 0.33-1.01, $p=0.042$). However these results did not sustain multiple testing corrections. A total of 12 promoter haplotypes (pHaps) were inferred but only haplotypes TGH*1, *2 and *3 had frequencies ≥ 0.05 and represented 98% of the observed haplotypes in the population tested (Table 2). The remaining 2% of the chromosomes carried nine minor haplotypes that were grouped under TGH*. We found no significant difference in the overall distribution of the 12 *TGFB1*-derived pHaps between cases and controls (Global $X^2=10.85$, 11 degrees of freedom, $p=0.46$) and found no evidence of association between individual haplotypes and the risk of pre-B ALL. Investigating multi-SNP combinations, the best result was obtained for marker combination -1886A/-1550AGG/-509C, which was associated with a

decreased risk of pre-B ALL (OR= 0.86) however the global test statistic failed to reach significance ($p=0.15$).

Though the expected variability of *TGFB1* expression levels due to promoter SNPs could indeed contribute to leukemogenesis, our data did not support a role for *TGFB1* promoter variants -1886G>A, -1571A>G, -1550DEL/AGG or -509C>T in the etiology of childhood pre-B ALL, at least among the French Canadian population. Having screened 2kb upstream of the transcription start site in 40 individuals (7), a strength of this study was our ability to capture and survey most common genetic variation within the region under investigation with reasonable statistical power. Despite the relatively small size of this dataset, we had 80% power at the 5% level to detect a minimum OR of 1.5 with a minor allele frequency $\geq 15\%$. However it remains possible that other variants within the *TGFB1* gene and surrounding regulatory sequences that were not assessed within the scope of this study could modify disease susceptibility. Additional analyses in independent datasets are required to further support the lack of association between these *TGFB1* variants and childhood pre-B ALL since our study did have limited statistical power to detect associations involving weak effects or rarer variants.

Tables

Table 1. Allele and genotype frequencies of pSNPs in *TGFB1* in childhood pre-B ALL patients and controls.

DNA variants, genotypes and alleles	N. (%)		OR (95% CI)	p
	ALL patients	Controls		
-1886G>A				
Alleles				
-1886G	215 (34.0)	201 (31.3)	1 (referent)	—
-1886A	417 (66.0)	441 (68.7)	0.88 (0.70-1.12)	0.30
Genotypes				
GG	40 (12.7)	25 (7.8)	1 (referent)	—
GA	135 (42.7)	151 (47.0)	0.56 (0.32-0.97)	0.038
AA	141 (44.6)	145 (45.2)	0.61 (0.35-1.05)	0.076
GA+AA vs. GG	276 (87.3)	296 (92.2)	0.58 (0.33-1.01)	0.042
-1571A>G				
Alleles				
-1517A	254 (41.1)	253 (39.4)	1 (referent)	—
-1517G	364 (58.9)	389 (60.6)	0.93 (0.74-1.17)	0.54
Genotypes				
AA	52 (16.8)	45 (14.0)	1 (referent)	—
AG	150 (48.5)	163 (50.8)	0.80 (0.50-1.26)	0.33
GG	107 (34.6)	113 (35.2)	0.82 (0.51-1.32)	0.42
-1550DEL/AGG				
Alleles				
-1550AGG	417 (67.0)	448 (69.1)	1 (referent)	—
-1550DEL/AGG	205 (33.0)	200 (30.9)	1.10 (0.87-1.39)	0.42
Genotypes				
AGG AGG	142 (45.7)	150 (46.3)	1 (referent)	—
AGG DEL/AGG	133 (42.8)	148 (45.7)	0.95 (0.68-1.32)	0.76
DEL/AGG DEL/AGG	36 (11.6)	26 (8.0)	1.46 (0.84-2.55)	0.18
-509C>T				
Alleles				
-508C	418 (66.6)	446 (68.8)	1 (referent)	—
-508T	210 (33.4)	202 (31.2)	1.11 (0.88-1.40)	0.39
Genotypes				
CC	141 (44.9)	148 (45.7)	1 (referent)	—
CT	136 (43.3)	150 (46.3)	0.95 (0.69-1.32)	0.77
TT	37 (11.8)	26 (8.0)	1.49 (0.86-2.59)	0.15

Percentages indicate number of individuals with a given genotype/total number of individuals in the dataset or the number of chromosomes with given allele/total number of chromosomes in the dataset. OR indicates crude odds ratio; —, not applicable.

Table 2. Distribution of *TGFB1* promoter haplotypes in pre-B ALL patients and controls.

Haplotype	DNA variant				N. (%)		OR (95% CI)	<i>p</i>	Global X ² (df)	Global <i>p</i>
	-1886G>A	-1571A>G	-1550DEL/AGG	-509C>T	ALL patients	Controls				
TGH*1	A	G	AGG	C	359 (56.45)	386 (59.75)	0.87 (0.69-1.10)	0.23	10.85 (11)	0.46
TGH*2	G	A	DEL/AGG	T	205 (32.24)	196 (30.34)	1.09 (0.86-1.39)	0.46		
TGH*3	A	A	AGG	C	50 (7.86)	53 (8.21)	0.95 (0.62-1.46)	0.82		
TGH*	*	*	*	*	22 (3.45)	11 (1.70)	—	—		

The risk of ALL was evaluated for each haplotype compared with all other possible haplotypes combined. Percentages indicate number of chromosomes with given haplotype/total number of chromosomes. Haplotypes with relative frequencies <5% are grouped under TGH* and are represented as * combinations of the four DNA variants. A likelihood ratio test was performed in FAMHAP to compare global haplotype differences between cases and controls and is reported here as a Global X² test with number of haplotype parameters different from zero-1 degrees of freedom. OR indicates crude odds ratio; df, degrees of freedom; and —, not applicable.

References

1. Elliott RL, Blobe GC. Role of transforming growth factor Beta in human cancer. *J Clin Oncol*. 2005 Mar 20;23(9):2078-93.
2. Dunning AM, Ellis PD, McBride S, Kirschenlohr HL, Healey CS, Kemp PR, et al. A transforming growth factor beta1 signal peptide variant increases secretion in vitro and is associated with increased incidence of invasive breast cancer. *Cancer Res*. 2003 May 15;63(10):2610-5.
3. Kang HG, Chae MH, Park JM, Kim EJ, Park JH, Kam S, et al. Polymorphisms in TGF-beta1 gene and the risk of lung cancer. *Lung Cancer*. 2006 Apr;52(1):1-7.
4. Saltzman BS, Yamamoto JF, Decker R, Yokochi L, Theriault AG, Vogt TM, et al. Association of genetic variation in the transforming growth factor beta-1 gene with serum levels and risk of colorectal neoplasia. *Cancer Res*. 2008 Feb 15;68(4):1236-44.
5. Ewart-Toland A, Chan JM, Yuan J, Balmain A, Ma J. A gain of function TGFB1 polymorphism may be associated with late stage prostate cancer. *Cancer Epidemiol Biomarkers Prev*. 2004 May;13(5):759-64.
6. Healy J, Belanger H, Beaulieu P, Lariviere M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood*. 2007 Jan 15;109(2):683-92.
7. Sinnett D, Beaulieu P, Belanger H, Lefebvre JF, Langlois S, Theberge MC, et al. Detection and characterization of DNA variants in the promoter regions of hundreds of human disease candidate genes. *Genomics*. 2006 Jun;87(6):704-10.
8. Labuda D, Krajcinovic M, Richer C, Skoll A, Sinnett H, Yotova V, et al. Rapid detection of CYP1A1, CYP2D6, and NAT variants by multiplex polymerase chain reaction and allele-specific oligonucleotide assay. *Anal Biochem*. 1999 Nov 1;275(1):84-92.
9. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet*. 1998 Jul;63(1):259-66.
10. Becker T, Knapp M. Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet Epidemiol*. 2004 Jul;27(1):21-32.
11. Becker T, Knapp M. A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am J Hum Genet*. 2004 Oct;75(4):561-70.

12. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B*. 1995;57(1):289-300.

Reprint Permissions

Sorry for answering so late. Please use this mail as a permission granted to your needs. Please note that at the moment you can only indicate the article published in July 2009. As regards the ms HAEMATOL/2010/022459, you cannot indicate it until it is not published. It will be probably published as E-Pub in the next weeks. If you agree, I will send you the details as soon as it is published.

To whom it may concern:

we herewith grant permission to re-use the article cited in your message below free of charge: this publication is intended for no-profit use. Any other use is discouraged and it will remain under copyright laws of the Ferrata Storti Foundation, Pavia, Italy. You are also obliged to correctly quote the original article reference (Healy J, Roy-Gagnon MH, Sinnett D. No evidence for association between TGFB1 promoter SNPs and the risk of childhood pre-B acute lymphoblastic leukemia among French Canadians. Haematologica 2009; 94 (7):1034-6).

Best regards

[Quoted text hidden]

Lorella Ripari
Haematologica Journal Office

Functional Impact of Sequence Variation in the Promoter Region of *TGFB1*

Jasmine Healy,^{1*} Joëlle Dionne^{1*}, Hélène Bélanger¹, Mathieu Larivière¹, Patrick Beaulieu¹, Damian Labuda,^{1,2} and Daniel Sinnett^{1,2}

¹Division of Hematology-Oncology, Sainte-Justine Hospital Research Center, University of Montreal, QC, Canada;

²Department of Pediatrics, Faculty of Medicine, University of Montreal, QC, Canada

* Jasmine Healy and Joëlle Dionne contributed equally to this work

This research was originally published in the *International Journal of Cancer*. Healy J, Dionne J, Bélanger H, Larivière M, Beaulieu P, Labuda D, Sinnett D. Functional impact of sequence variation in the promoter region of *TGFB1*. Int J Cancer. 2009 Sep 15;125(6):1483-9. © Wiley-Liss Inc., A Wiley Company. The article is included with the permission of the authors and the journal.

Author Contributions

For this article I performed 50% of the work, including writing the manuscript. J. Dionne, H. Bélanger and M. Larivière performed the molecular genetics experiments and functional validation. P. Beaulieu performed *in silico* analysis and D. Labuda was involved in the world population data collection and haplotype reconstruction. I gathered, analyzed and interpreted the results and wrote the article. D. Sinnett is principle investigator and participated in the design and coordination of the study as well as writing the manuscript. All authors read and approved the final manuscript.

Abstract

Pathological deregulation of the transforming growth factor, beta 1 (TGFB1) pathway has been implicated in the development of several major diseases, including cancers. Regulatory variation in the TGFB1 gene may lead to altered *TGFB1* expression and activity, and thus, modulate an individual's susceptibility to disease. Here, we performed a study of the functional relevance of *cis*-acting regulatory variation in the proximal promoter region of the *TGFB1* gene. In a previous study, 9 promoter polymorphisms were identified in the 2kb region upstream of the transcription start site and 9 distinct promoter haplotypes were inferred from a panel of individuals from 5 distinct continental population groups. Following experimental validation, we found that the 2 major haplotypes significantly influenced *TGFB1* transcriptional activity in an allele-specific manner and that 3 of the SNPs (–1886G>A, –509C>T and –1550DEL/AGG) altered DNA-protein complex formation. Though the biological relevance of these findings remains to be verified, our study suggests that polymorphisms in the *TGFB1* promoter could indeed influence gene expression and potentially contribute to the pathogenesis of *TGFB1* related diseases.

Introduction

Transforming growth factor, beta 1 (TGFB1) is a potent growth inhibitor to a wide variety of cell types, and plays a central role in regulating passage through the G1/S checkpoint. As a multifunctional cytokine, TGFB1 regulates a number of important cellular responses in addition to its role in cell cycle control including cell differentiation, apoptosis, cell migration, immune responses and angiogenesis (1, 2). Pathological deregulation of the TGFB1 signaling pathway has been extensively studied and has been shown to be implicated in the development of several major disease groups, including cancers (3), atherosclerosis (4), fibrotic diseases (5), and obesity (6). Its involvement in carcinogenesis is complex, having both tumour suppressor and oncogenic activities (7). A stage specific duality of function is the emerging paradigm for the role of *TGFB1* in cancer (2): mutational inactivation or altered expression of the gene appears to be important in the early stages of cancer development in order for tumor cells to escape from its antiproliferative effects (8); whereas *TGFB1* overexpression would be necessary in the later stages to enhance tumor progression and increase metastasis (9). Dose-dependent changes in gene expression of *TGFB1*, and of other downstream members of the pathway, are sufficient to disrupt signaling and promote tumorigenesis (7). Accordingly, DNA variation in gene promoter and/or regulatory sequences could impact *TGFB1* expression levels and potentially play a role in disease etiology. In particular, one of the more commonly studied *TGFB1* SNPs, -509C>T (rs1800469) located in the proximal promoter region has been associated with variability in TGFB1 plasma levels and modified risk of breast (10, 11), lung (12), colorectal (13, 14), and prostate (15) cancers. Here, we performed a comprehensive study of *cis*-acting regulatory variation in the *TGFB1* promoter. Using *in silico* and *in vitro* approaches, we investigated the effects of *TGFB1* promoter SNPs (pSNPs) -1960C>T (rs11466311), -1886G>A (rs2317130), -1571A>G (rs4803457), -1550DEL/AGG (rs11466313), -1349INS/C

(rs36204367), -1146C>G (rs3087453), -800G>A (rs1800468), -509C>T (rs1800469), -448C>T (rs11466314) and 2kb *TGFB1* promoter haplotypes (pHaps) on differential DNA-protein binding, and promoter activity. We showed that polymorphisms in the *TGFB1* promoter can disrupt protein binding, and influence promoter activity in an allele-specific manner, suggesting that *TGFB1* promoter variation may be important in disease.

Material and Methods

Promoter SNP identification

In a previous study, we screened the 2kb sequence upstream of the *TGFB1* transcription start site for polymorphisms on a population panel consisting of 39 unrelated individuals from 5 continental groups (7 Africans, 8 Europeans, 8 South East Asians, 8 Middle Easterners and 8 Amerindians) as well as a chimpanzee sample, obtained from the Granby Zoo (Granby, Canada), to infer ancestral alleles (refer Sinnott et al. (16) and Labuda et al. (17) for details). The targeted proximal promoter region of the *TGFB1* gene was amplified sequentially in 7 overlapping fragments of ~350 bp for dHPLC analysis followed by direct sequencing. Primer sequences and detailed information on experimental conditions are available upon request. Polymorphism positions relative to the transcription start site were identified based on the RefSeq mRNA sequence NM_000660.1 with nucleotide +1 at position chr19:46,551,656 of the March 2006 human reference sequence (NCBI Build 36.1).

In silico analysis

Computational haplotype reconstruction was performed using the PHASE software (Version 2.1) (18) and frequencies were estimated for each of the 5 population groups in Labuda et al. (17). *In silico* analysis of the *TGFB1* gene promoter sequence was performed using the MatInspector software (19) to predict the presence of putative transcription factor binding sites (TFBSs). For each pSNP under study, 50 bp of surrounding genomic sequence (25 bp upstream and 25 bp downstream of both the major and minor alleles) were analyzed using the optimized matrix similarity thresholds to predict gain and/or loss of a putative TFBS.

Gene reporter assays

Constructs. Four *TGFB1* pHaps (TGH1, TGH2, TGH3 and TGH4) were selected for reporter gene assays. For the higher-frequency haplotypes (TGH1 and TGH2), genomic DNA of known heterozygous individuals were PCR amplified to obtain 2kb haplotype-specific fragments that were then subcloned into the promoterless pGL3-Basic *Firefly* luciferase reporter vector (Promega, Madison, WI). The lower-frequency haplotypes (TGH3 and TGH4) were generated by site-directed mutagenesis (Stratagen, Cedar Creek, TX) using TGH1 as template, following the manufacturer's instructions. The oligonucleotides used for site-directed mutagenesis were purchased from Integrated DNA Technologies (Coralville, IA). Sequences are available upon request. The resulting constructs were sequenced to confirm the presence of the expected polymorphic sites and then purified on QIAquick PCR purification columns (Qiagen, Mississauga, Ontario) prior to transfection.

Transient transfections. Reporter gene expression was measured transiently in 3 human cell lines: HepG2 (hepatoma), Jeg-3 (choriocarcinoma), HeLa (cervical

carcinoma) cells. For each, $\sim 6\text{--}8 \times 10^3$ cells were plated-out and grown in 96-well plates (30 mm^2) to reach 80–90% confluence at the time of transfection. Cells were transfected with lipofectamine according to the manufacturer's protocol (Invitrogen, Burlington, Ontario). All 3 cell lines were co-transfected with 100 ng of each haplotype-specific construct and 0.5 ng of a CMV-driven *Renilla* luciferase pRL-CMV plasmid (Promega, Madison, WI) (ratio 200:1) to control for transfection efficiency. Similar experiments were performed with a negative control consisting of the empty promoterless pGL3-Basic plasmid (Promega, Madison, WI) and also with a SV40-driven *Firefly* luciferase pGL3-Control plasmid (Promega, Madison, WI) as a positive control. Cells were harvested 24hrs post-transfection and luciferase reporter gene activities were measured with the Dual-Luciferase Reporter Assay System according to the manufacturer's instructions (Promega, Madison, WI), in a Spectra Max 190 luminometer (Molecular Devices, Sunnyvale, CA). The *Renilla* luciferase activity of the co-transfected pRL-CMV internal control was used to normalize the results of the *Firefly* luciferase activity. Results are expressed as the ratio of *Firefly* luciferase activity divided by the pRL-CMV *Renilla* activity and are represented as the relative luciferase activity of 4 independent replicates (Mean \pm SD). The promoterless pGL3-basic vector was used to measure basal expression levels in each cell line. Luciferase levels for promoter haplotypes TGH1, TGH2, TGH3 and TGH4 were normalized against the basal pGL3-basic expression level in each cell line to standardize results and help to differentiate any intrinsic cell-specific effects on promoter expression levels (Figure 2). A significant 1.5-fold difference in luciferase activity ($p \leq 0.05$) over the empty vector was considered evidence for haplotype-specific promoter activity. Pairs of haplotypes were compared by Student's unpaired *t*-test using promoter haplotype TGH1 as reference against which relative expression was compared.

Electrophoretic mobility shift assays

Electrophoretic mobility shift assays (EMSAs) were performed on the 4 most frequent pSNPs (–1886G>A, –1571A>G, –1550DEL/AGG and –509C>T). For each pSNP, doublestranded oligonucleotide probes corresponding to the sequences surrounding the polymorphic sites were radiolabeled and allowed to interact with nuclear extracts prepared from HepG2, Jeg-3 and HeLa cells as described in Belanger et al. (20), with the exception that, for binding competition, the EMSA was performed in the presence of 50-fold molar excess of the unlabeled probe oligonucleotide, the unlabeled corresponding mutant probe or 1 lg of calf thymus DNA (for variants –1886G>A, –1571A>G and –509C>T) or a random oligonucleotide (–1550DEL/AGG) as non-specific competitor. The oligonucleotides used in the electrophoretic mobility shift assays are as follows (only the 5'-3' top strand of the double-stranded oligonucleotides are shown):

–1886G>A, AATTACCACCATCT(G/A)ATCTACCCTATA;
 –1571A>G, GAAAGAAGGCCCTGGGCCCAAAG(A/G)GAGCAGGGCAG;
 –1550DEL/AGG, GGGACATGAGG(AGG/-)GAAGGCAG;
 –509C>T, CTGACCCTTCCATCC(C/T)TCAGGTGTCCTGT;
 Random oligonucleotide, GATCGAACTGACCGCTTGCGGCCCCGT.

Results

The haplotype distributions among the 5 continental population groups are shown in Table 1 and Figure 1. (Adapted from Labuda et al. (17)). Haplotype structure analysis (refer (17)) showed that the 4 most frequent pSNPs (–1886G>A, –1571A>G, –1550DEL/AGG and –509C>T) split the *TGFB1* haplotypes into 2 clusters centered around the 2 major pHaps that together accounted for 87.2% of the total haplotypic diversity: TGH1 (46.2%) carrying

pSNPs -1886G>A and -1571A>G, and TGH2 (41.0%) carrying pSNPs -1550DEL/AGG and -509C>T. A third branch derived from the ancestral haplotype sequence consisted of African-specific polymorphisms only.

To further assess the functional significance of DNA variation in the promoter region of *TGFB1*, we performed promoter activity analysis using luciferase gene reporter assays. Because of the observed inter-population differences in haplotypic diversity and frequency, we limited our analysis to mainly European-specific haplotypes, that is haplotypes TGH1, TGH2, TGH3 and TGH4. For each of these haplotypes, the whole 2kb promoter sequence was subcloned in the promoterless pGL3-Basic *Firefly* luciferase reporter vector. The ability of each sequence to promote transcription of the luciferase gene was tested transiently in 3 human cell lines (HepG2, Jeg-3 and HeLa). As shown in Figure 2, the HeLa cell line had overall lower expression levels (relative luciferase activity ranging from 0.3 to 1.8, following normalization with the empty pGL3-Basic vector) compared to Jeg-3 and HepG2. HepG2 showed overall highest luciferase expression from the haplotype specific promoter constructs with relative luciferase activity ranging from 2.9 to 11.9. Promoter haplotype TGH1 showed weak promoter activity in HeLa (1.7-fold over the empty vector), moderate promoter activity in Jeg-3 (3-fold over the empty vector) and strong activity in HepG2 (12-fold over the empty vector). Similarly, haplotypes TGH3 and TGH4 showed intermediate to high promoter activity in Jeg-3 and HepG2 but weak activity in HeLa. Haplotype TGH2 showed overall weaker promoter activity in all 3 cell lines.

We compared each haplotype to the relative promoter activity of TGH1. Significant differences were found between the 2 major haplotypes, TGH1 and TGH2, across all 3 cell lines (Figure 2). Relative luciferase activity driven by TGH2 was 1.5-fold to 5-fold lower than the luciferase levels driven by its

counterpart TGH1, indicating allele-specific promoter activity. Comparison of pHaps TGH1 and TGH2 indicates that either the deletion at position –1550 or the variant T allele at pSNP –509 or the ancestral G and A alleles at positions –1886 and –1571, respectively, or a combination of these 4 genotypes, lowers the promoter activity by over 1.5-fold. On the contrary, the opposite alleles carried on TGH1 may act to increase promoter activity. Haplotypes TGH3 and TGH4 showed significant increased and decreased promoter activity, respectively, when compared to TGH1, in at least one of the cell lines tested, however, these differences were not consistent across cell lines.

To further investigate the alleged effect of *TGFB1* promoter variation on gene expression, we looked at the putative impact pSNPs could have on DNA-protein binding capacity. Depending on its position in a given regulatory region, a pSNP could potentially disrupt transcription factor recognition and thereby affect gene expression. The 4 most frequent variants, –1886G>A, –1571A>G, –1550DEL/AGG and –509C>T, were examined. Using *in silico* analysis, we first assessed the possibility that the pSNPs laid in consensus sequence elements for transcription factor recognition. Using the MatInspector software, we found that all 4 pSNPs had a predicted impact on TFBS motifs and that all 4 led to both the loss and/or gain of putative binding sites (Table 2), supporting a functional role for these pSNPs. These observations suggested that the haplotype-specific differences observed in the gene reporter assays could in fact be the consequence of polymorphic *cis*-elements in the promoter.

In an attempt to identify the specific variants responsible for the changes in expression, pSNPs –1886G>A, –1571A>G, –1550DEL/AGG and –509C>T were further validated by EMSAs in 3 different cell lines. Representative data in Jeg-3 are shown in Figure 3. For variant –1886G>A, we observed 3 distinct slow-migrating DNA-protein complexes. Allele-specific differential binding was

observed for Complex 3 which showed high binding affinity for the –1886A allele (Figure 3, lane 6) but was hardly evident when incubated with Jeg-3 nuclear extract in the presence of the –1886G allele (Figure 3, lane 1). To confirm binding specificity to the –1886A allele, competition EMSAs were performed. In the presence of a 50-fold molar excess of the unlabeled –1886G probe, Complex 3 was still evident albeit at a much lower level, whereas the complex almost completely disappeared when competed with a 50-fold excess of the corresponding unlabeled –1886A allele (Figure 3, lane 7 vs. lane 8). The nuclear protein bound to the –1886A allele in Complex 3 appeared therefore to be a specific DNA-protein interaction that was weakened in the presence of the –1886G allele. We may speculate that the –1886G allele, found on the low expressing TGH2 haplotype may hinder protein binding within this region of the promoter, which in the context of transcription factor binding, could lead to decreased promoter activity, as observed in the luciferase gene reporter assay. Similar results were obtained for the –1886G>A variant with HeLa and HepG2 nuclear extracts (Table 2), which both showed similar allele-specific complex formation. Promoter SNPs –1571A>G, –1550DEL/AGG and –509C>T were processed similarly and we found that –1550DEL/AGG also showed differential allelic shifts in all 3 cell lines tested (refer results for Jeg-3 in Figure 3), as did –509C>T, however, in HepG2 cells only (Table 2). The –509T allele disrupted complex formation in HepG2 cells also suggesting a potential implication for this variant in the observed differences in reporter gene expression, whereas in the presence of the –1550DEL/AGG allele, allele DNA-protein complexes were gained (Figure 3) suggesting that the deletion at position –1550 may lead, on the contrary, to novel TFBS motifs.

Discussion

TGFB1 is a multifunctional cytokine that is ubiquitously expressed and plays a crucial role in normal cellular processes and disease. The TGFB1 signaling cascade is highly regulated and is sensitive to protein level variation (7). DNA variants that lie in putative TFBSs in the *TGFB1* promoter could therefore potentially perturb transcription factor binding, alter gene expression levels, and have a number of pathological consequences. In this study, we set out to investigate the impact of functional diversity in the promoter region of *TGFB1* on the transcriptional regulation of gene expression. Nine pSNPs were previously identified in the proximal promoter region of the *TGFB1* gene and haplotype inference revealed a total of 9 distinct promoter haplotypes among 5 continental populations (17). Two major haplotypes, TGH1 and TGH2 represented, respectively, 46.2% and 41.0% of the observed worldwide haplotypic variability. Allele combinations –1886A/–1571G and –1550DEL/AGG/–509T were mutually exclusive, creating 2 distinct clusters around haplotypes TGH1 and TGH2. The ancestral promoter haplotype (“ANC” in Figure 1) derived from chimpanzee DNA, was located between these 2 clusters. A third branch derived from the ancestral haplotype sequence consisted of African-specific polymorphisms only (refer haplotype network, Figure 1).

These clusters could represent distinct phenotypic groups. Our *in vitro* functional assays supported this hypothesis. Using luciferase gene reporter assays and electrophoretic mobility shift assays, we provided evidence of allele-specific differences in promoter activity and DNA-protein complex formation. Three different human cell lines were used in these studies, HeLa (cervical carcinoma), Jeg-3 (choriocarcinoma) and HepG2 (hepatoma) cells, to assess tissue-specific effects. As expected, not all observations were consistent across cell lines, highlighting the importance of *in vivo* validation to determine the

biological relevance of these results. The luciferase gene reporter assays did, however, provide evidence of differential promoter activity between TGH1 and TGH2 that was consistent across all 3 cell lines. We observed decreased expression levels for pHap TGH2 when compared to TGH1. A previous report by Shah et al. also assessed diversity in the *TGFB1* promoter region and reported similar results (21). While the authors examined an extended regulatory region which included exon 1, the haplotype patterns and frequencies from both studies are comparable (TGH1 vs. p003 and TGH2 vs. p001 of the Shah et al. study) and similarly, distinct phylogenetic clusters were identified for the 2 most common haplotypes p003 and p001. Interestingly, Shah et al. also examined multiple samples of various racial backgrounds and identified a number of novel African-American specific polymorphisms including SNPs –448C>T and –1146C>G that were observed uniquely among African individuals in our study. Though the impact of these variants on gene expression among this population was not investigated within the scope of this study, Shah et al. showed that variant –448C>T alters the binding affinity of 2 unidentified transcription factor complexes which also translates into a significant difference in reporter gene expression (21).

Our EMSAs showed that variants –1886G>A, –1550DEL/AGG and –509C>T perturbed protein binding. TGH2 alleles –1886G and –509T lead to the loss of specific DNA-protein complex formation while the AGG deletion lead to the gain of protein complex binding at position –1550. These results were corroborated by the *in silico* analysis, which predicted loss and gains of putative TFBS motifs at loci –1886G>A, –1550DEL/AGG and –509C>T. Variant –1886G>A was shown to lead to the predicted loss of 3 putative TFBSs, AP4, NeuroD/BETA2 and GATA binding factors. Transcription factor AP4 acts as part of a repressor complex to negatively regulate expression of target genes and has been shown to directly repress expression of cyclin dependent kinase inhibitor 1A (CDKN1A) by occupying 4 CAGCTG motifs in the CDKN1A promoter and was shown to

interfere with cell cycle arrest during monoblast differentiation (22).

NeuroD/BETA2, a basic helix-loop-helix transcription factor, has been shown to play a role in tissue-specific differentiation of pancreatic and enteroendocrine cells and regulates insulin gene expression and β -cell differentiation (23), while GATA transcription factors play a major role in regulating hematopoiesis and have been involved in many hematologic disorders (24). Variant -509C>T abrogated a putative olfactory neuron-specific transacting factor involved in the cell-specific expression of olfactory marker protein genes. The 3 bp deletion at position -1550 was predicted to lead to the loss of PAX and MAZ transcription factors. Both PAX4 and PAX6 are expressed in pancreatic islets and, along with NeuroD/BETA2, regulate endocrine fate in the pancreas (25) whereas Myc associated zinc finger transcription factors are speculated to have dual roles in transcription initiation and termination and have been shown to be expressed in all tissues except the kidney (26). However, the gel shift analysis showed appearance, rather than loss, of protein complexes in the presence of the deletion. The only transcription factor that was predicted to be gained by the -1550AGG deletion was the chorion-specific transcription factor GCMA, involved in the expression of multiple placenta-specific genes (27). Although the biological relevance of these predictions needs to be further investigated in the context of disease, the putative loss and/or gain of transcription factor binding motifs did translate into significant differences in reporter gene expression and disruption of DNA-protein complex formation. And though the identity of these protein complexes remains to be verified, it is possible that *TGFB1* alleles -1886G, -1550DEL/AGG and -509T could contribute to the decreased expression levels observed for haplotype TGH2.

This study was limited to the proximal promoter region (2kb region upstream of the transcription start site), and it is thus likely that other regulatory SNPs were missed. In this regard, previous studies characterizing the promoter region of *TGFB1* identified sequences responsible for both promoting and inhibiting

transcriptional activity within the 3.1kb region upstream of the transcription start site and the first exon of the *TGFB1* gene. An additional 9 polymorphisms were also identified within this extended region (21). According to these studies, variant –509C>T appeared to reside in a negative regulatory region (28) whereas variants –1550DEL/AGG and –1571A>G, and perhaps –1886G>A, were shown to lie in a distal enhancer region that dramatically increased *TGFB1* expression (21). Moreover, pSNPs –800G>A (rs1800466) and –448C>T (rs11466314), which were not thoroughly investigated here, have previously been associated with variations in gene expression, as have variants –387C>T (rs11466315) in the 5' UTR, SNP rs11466345 at the 3' end of *TGFB1* and the Leu10Pro substitution at codon 10 (rs1800470), which were outside of the scope of this study (5, 13, 21, 29). Despite this limitation, it appears clear, by our study as well as previous reports, that regulatory variation within and around the *TGFB1* sequence, can modulate transcriptional activity and potentially have important implications in disease.

The –509C>T pSNP has been linked to inter-individual variations in *TGFB1* plasma levels to the extent where individuals that were homozygous for the –509T variant had almost double the level of circulating *TGFB1* compared to –509CC carriers (30). A number of studies investigating the association of –509C>T with numerous diseases have been performed, however, they have yielded largely mixed results. Shah et al. were able to demonstrate exclusive *in vivo* and *in vitro* recruitment of AP1 to –509C leading to transcriptional repression of the *TGFB1* gene. However, our *in silico* analysis failed to predict the gain and/or loss of this particular transcription factor binding motif at position –509 and we found that haplotype TGH2 which carries the –509T allele leads, on the contrary, to weaker promoter activity and decreased expression levels. This discrepancy reflects the limitations of computational tools that rely on prior knowledge for modeling and predicting biological processes, once again highlighting the importance of experimental validation. These discrepancies

could also be indicative of a haplotypic effect observed in our study, in which the combined effects on gene expression of a number of promoter sequence variants were accounted for, rather than focusing on a single polymorphism.

The mechanisms that control *TGFB1* expression remain unclear and there still remains much to be learned about the genetic and molecular implications of *TGFB1* in disease before it can be used as a marker in prognosis and prevention. For example, the functional impact of many of the SNPs identified in this study, as well as others in extended regulatory regions need to be further explored and the population-specific effects of *TGFB1* regulation warrant more ample investigation. This is especially true since a number of promoter SNPs investigated here and in previous reports appear to be African-specific. Though variation in *TGFB1* plasma levels has been reported in a number of diseases including cancers, the role of *TGFB1* promoter variation in modulating expression and in the modulation of disease risk remains unclear. Overall, this study should help further elucidate the impact of functional diversity in the promoter region of *TGFB1* on transcriptional regulation and gene expression and be useful for future studies to determine the value of these *TGFB1* promoter variants in the clinical setting.

Figures

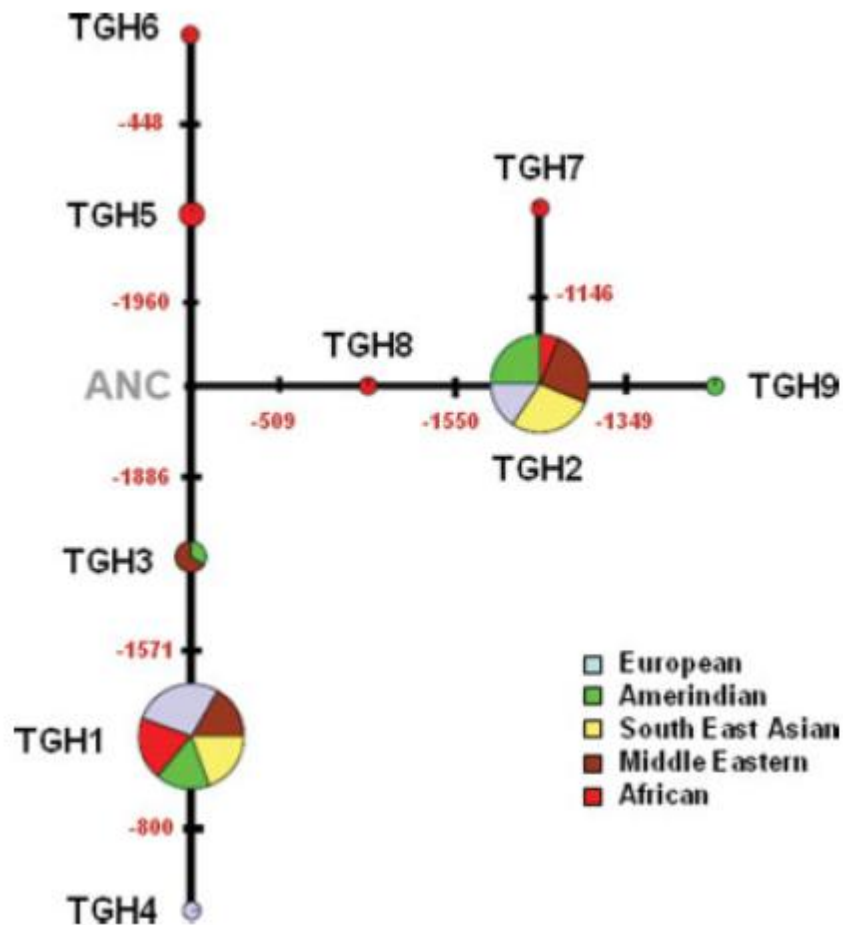


Figure 1. *TGFβ1* haplotype network

This haplotype network was adapted from Labuda et al. (17). The area of the circle is proportional to the overall haplotype frequency, while colors indicate the distribution among continental groups. The solid lines connecting the haplotypes represent single mutations occurring without recombination, and correspond to the maximum parsimony tree for this network. “ANC” designates the ancestral promoter haplotype derived from chimpanzee DNA (refer to Labuda et al. for details (17)).

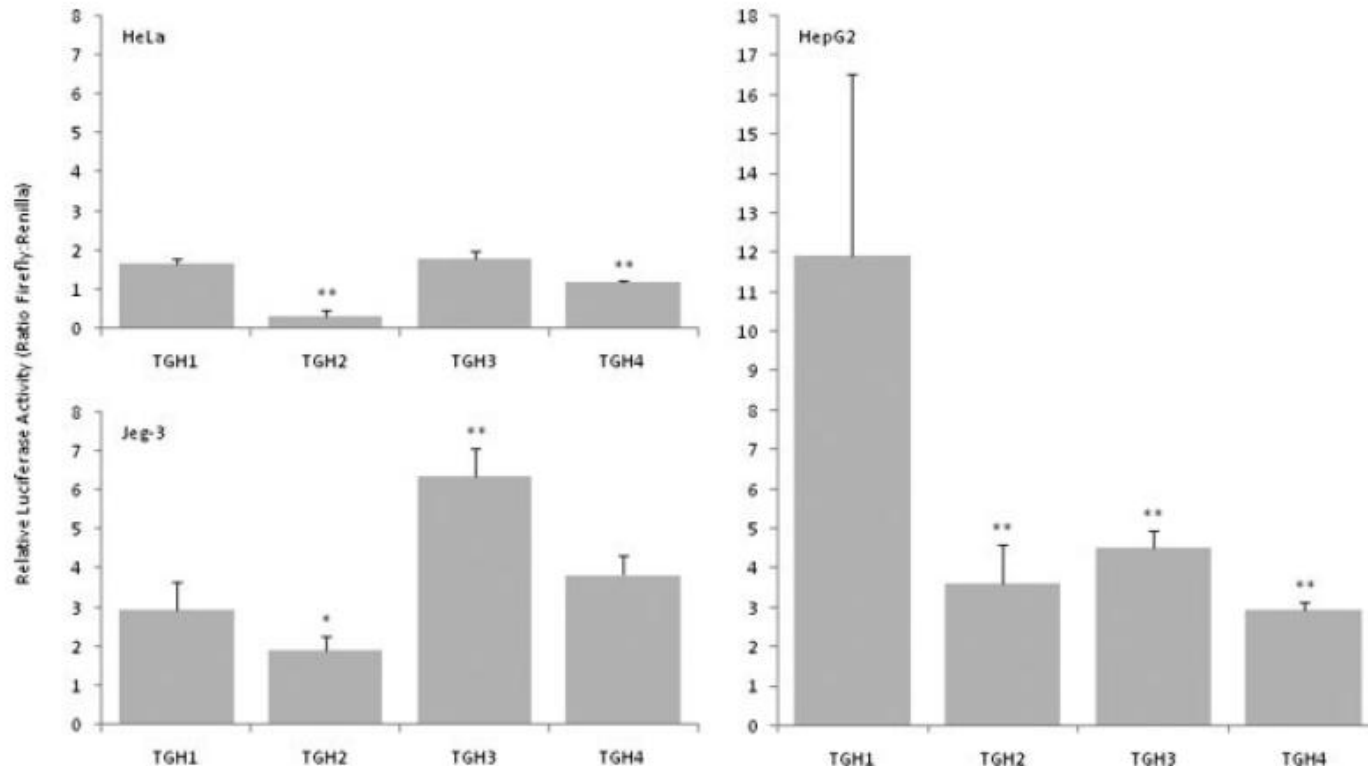


Figure 2. Gene reporter assays evaluating promoter activity of 4 promoter haplotypes of *TGFB1*

Relative luciferase activity (Mean \pm SD) of *TGFB1* promoter haplotypes was measured following transient transfection in HeLa, Jeg-3 and HepG2 cells. The ratio of *Firefly: Renilla* activity of each promoter was normalized against that of the empty promoterless pGL3-basic vector in each cell line. Promoter haplotype TGH1 was used as reference against which pairwise comparisons were made. Significant differences between haplotype expression are marked with an asterisk (* $p < 0.05$; ** $p < 0.01$).

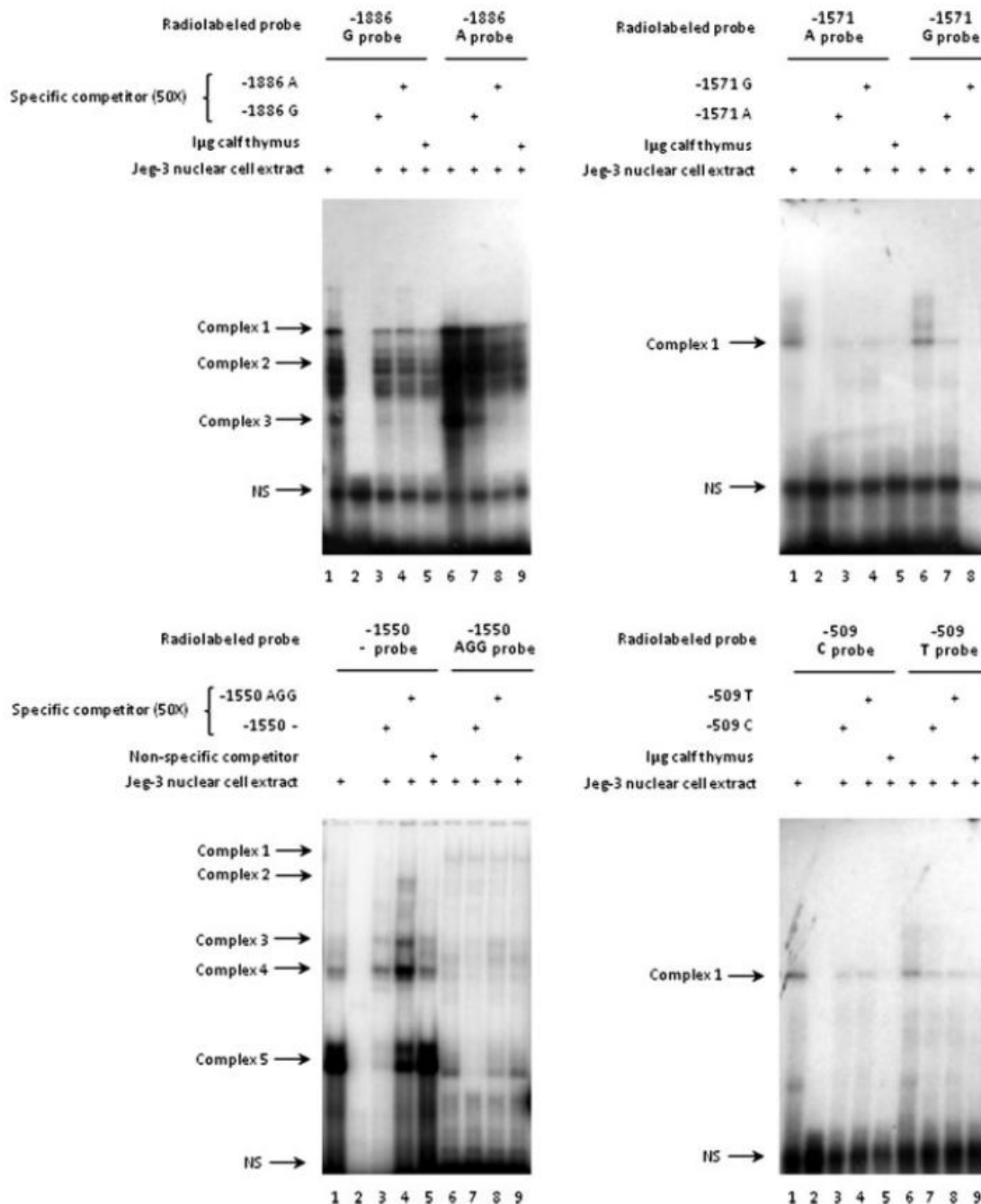


Figure 3. Representative EMSA analyses showing allelic DNA-protein interactions for pSNPs -1886G>A, -1571A>G, -1550DEL/AGG and -509C>T in Jeg-3

The unlabeled probes used to compete DNA-protein interactions are indicated (+) at the top of each lane. Specific competitors corresponding to the unlabeled allele-specific probes were used in lanes 3, 4, 7 and 8 and non-specific calf

thymus DNA or a random oligonucleotide was used in lanes 5 and 9. The position of the DNA-protein complexes of slower mobility are marked by arrows and fast migrating unbound probes can be seen at the very bottom of the gel. NS refers to non-specific complexes.

Tables

Table 1. *TGFB1* promoter haplotype distributions among five continental population groups

Haplotype	DNA-variant									Continental group					Total
	–1960	–1886	–1571	–1550	–1349	–1146	–800	–509	–448	European	Amerindian	South East Asian	Middle Eastern	African	
Ancestral:	C	G	A	AGG	–	C	G	C	C						
TGH1	C	A	G	AGG	–	C	G	C	C	10	6	7	6	7	36
TGH2	C	G	A	DEL/AGG	–	C	G	T	C	5	8	9	8	2	32
TGH3	C	A	A	AGG	–	C	G	C	C	–	1	–	2	–	3
TGH4	C	A	G	AGG	–	C	A	C	C	1	–	–	–	–	1
TGH5	T	G	A	AGG	–	C	G	C	C	–	–	–	–	2	2
TGH6	T	G	A	AGG	–	C	G	C	T	–	–	–	–	1	1
TGH7	C	G	A	DEL/AGG	–	G	G	T	C	–	–	–	–	1	1
TGH8	C	G	A	AGG	–	C	G	T	C	–	–	–	–	1	1
TGH9	C	G	A	DEL/AGG	INS/C	C	G	T	C	–	1	–	–	–	1

A total of nine promoter haplotypes (TGH1 to 9) were inferred from the continental populations under study. The ancestral haplotype was derived from chimpanzee DNA and used as reference to identify polymorphic alleles, shown here in bold. –, none.

Table 2. Impact of pSNPs on predicted transcription factor binding sites and summary of EMSA results

pSNP	<i>In silico predictions for TF binding sites</i>		<u>Validated differential binding</u>		
	Gain (Score) ¹	Loss (Score)	Hela	Jeg-3	HepG2
–886G>A	Phox2a (ARIX) and phox2b (0.89)	AP4 and related proteins (0.96) NeuroD, Beta2, HLH domain (0.95) GATA binding factors (0.91)	Yes	Yes	Yes
–1571A>G	Nuclear receptor subfamily 2 factors (0.82) Neuron-specific-olfactory factor (0.92)	Gut-enriched Krueppel like binding factor (0.90)	No	No	No
–1550DEL/AGG	Chorion-specific transcription factors with a GCM DNA binding domain (0.93)	PAX-4/PAX-6 paired domain binding sites (0.86) Myc associated zinc fingers (0.91)	Yes	Yes	Yes
–509C>T	HOX-PBX complexes (0.82) Gut-enriched Krueppel like binding factor (0.95)	Neuron-specific-olfactory factor (0.90)	No	No	Yes

Predicted impact of the pSNPs was estimated with the MatInspector software (Matrix Family Library Version 7.0). Gain and loss of transcription factor binding sites are specified for the variant allele with respect to the ancestral allele. Differential protein binding was further validated using electrophoretic mobility shift assays in three different cell lines.

¹Matrix similarity score.

References

1. Elliott RL, Blobe GC. Role of transforming growth factor Beta in human cancer. *J Clin Oncol*. 2005 Mar 20;23(9):2078-93.
2. Glick AB. TGFbeta1, back to the future: revisiting its role as a transforming growth factor. *Cancer Biol Ther*. 2004 Mar;3(3):276-83.
3. Shirai Y, Kawata S, Ito N, Tamura S, Takaishi K, Kiso S, et al. Elevated levels of plasma transforming growth factor-beta in patients with hepatocellular carcinoma. *Jpn J Cancer Res*. 1992 Jul;83(7):676-9.
4. Grainger DJ, Metcalfe JC. Transforming growth factor-beta: the key to understanding lipoprotein(a)? *Curr Opin Lipidol*. 1995 Apr;6(2):81-5.
5. Awad MR, El-Gamel A, Hasleton P, Turner DM, Sinnott PJ, Hutchinson IV. Genotypic variation in the transforming growth factor-beta1 gene: association with transforming growth factor-beta1 production, fibrotic lung disease, and graft fibrosis after lung transplantation. *Transplantation*. 1998 Oct 27;66(8):1014-20.
6. Long JR, Liu PY, Liu YJ, Lu Y, Xiong DH, Elze L, et al. APOE and TGF-beta1 genes are associated with obesity phenotypes. *J Med Genet*. 2003 Dec;40(12):918-24.
7. de Caestecker MP, Piek E, Roberts AB. Role of transforming growth factor-beta signaling in cancer. *J Natl Cancer Inst*. 2000 Sep 6;92(17):1388-402.
8. Markowitz SD, Roberts AB. Tumor suppressor activity of the TGF-beta pathway in human cancers. *Cytokine Growth Factor Rev*. 1996 Jun;7(1):93-102.
9. Reiss M. TGF-beta and cancer. *Microbes Infect*. 1999 Dec;1(15):1327-47.
10. Cox DG, Penney K, Guo Q, Hankinson SE, Hunter DJ. TGFB1 and TGFBR1 polymorphisms and breast cancer risk in the Nurses' Health Study. *BMC Cancer*. 2007;7:175.
11. Dunning AM, Ellis PD, McBride S, Kirschenlohr HL, Healey CS, Kemp PR, et al. A transforming growth factorbeta1 signal peptide variant increases secretion in vitro and is associated with increased incidence of invasive breast cancer. *Cancer Res*. 2003 May 15;63(10):2610-5.
12. Kang HG, Chae MH, Park JM, Kim EJ, Park JH, Kam S, et al. Polymorphisms in TGF-beta1 gene and the risk of lung cancer. *Lung Cancer*. 2006 Apr;52(1):1-7.

13. Saltzman BS, Yamamoto JF, Decker R, Yokochi L, Theriault AG, Vogt TM, et al. Association of genetic variation in the transforming growth factor beta-1 gene with serum levels and risk of colorectal neoplasia. *Cancer Res.* 2008 Feb 15;68(4):1236-44.
14. Chung SJ, Kim JS, Jung HC, Song IS. Transforming growth factor-[beta]1 -509T reduces risk of colorectal cancer, but not adenoma in Koreans. *Cancer Sci.* 2007 Mar;98(3):401-4.
15. Ewart-Toland A, Chan JM, Yuan J, Balmain A, Ma J. A gain of function TGFB1 polymorphism may be associated with late stage prostate cancer. *Cancer Epidemiol Biomarkers Prev.* 2004 May;13(5):759-64.
16. Sinnett D, Beaulieu P, Belanger H, Lefebvre JF, Langlois S, Theberge MC, et al. Detection and characterization of DNA variants in the promoter regions of hundreds of human disease candidate genes. *Genomics.* 2006 Jun;87(6):704-10.
17. Labuda D, Labbe C, Langlois S, Lefebvre JF, Freytag V, Moreau C, et al. Patterns of variation in DNA segments upstream of transcription start sites. *Hum Mutat.* 2007 May;28(5):441-50.
18. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 2001 Apr;68(4):978-89.
19. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, et al. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics.* 2005 Jul 1;21(13):2933-42.
20. Belanger H, Beaulieu P, Moreau C, Labuda D, Hudson TJ, Sinnett D. Functional promoter SNPs in cell cycle checkpoint genes. *Hum Mol Genet.* 2005 Sep 15;14(18):2641-8.
21. Shah R, Rahaman B, Hurley CK, Posch PE. Allelic diversity in the TGFB1 regulatory region: characterization of novel functional single nucleotide polymorphisms. *Hum Genet.* 2006 Mar;119(1-2):61-74.
22. Jung P, Menssen A, Mayr D, Hermeking H. AP4 encodes a c-MYC-inducible repressor of p21. *Proc Natl Acad Sci U S A.* 2008 Sep 30;105(39):15046-51.
23. Iwata I, Nagafuchi S, Nakashima H, Kondo S, Koga T, Yokogawa Y, et al. Association of polymorphism in the NeuroD/BETA2 gene with type 1 diabetes in the Japanese. *Diabetes.* 1999 Feb;48(2):416-9.
24. Harigae H. GATA transcription factors and hematological diseases. *Tohoku J Exp Med.* 2006 Sep;210(1):1-9.

25. St-Onge L, Sosa-Pineda B, Chowdhury K, Mansouri A, Gruss P. Pax6 is required for differentiation of glucagon-producing alpha-cells in mouse pancreas. *Nature*. 1997 May 22;387(6631):406-9.
26. Bossone SA, Asselin C, Patel AJ, Marcu KB. MAZ, a zinc finger protein, binds to c-MYC and C2 gene sequences regulating transcriptional initiation and termination. *Proc Natl Acad Sci U S A*. 1992 Aug 15;89(16):7452-6.
27. Yamada K, Ogawa H, Honda S, Harada N, Okazaki T. A GCM motif protein is involved in placenta-specific expression of human aromatase gene. *J Biol Chem*. 1999 Nov 5;274(45):32279-86.
28. Kim SJ, Glick A, Sporn MB, Roberts AB. Characterization of the promoter region of the human transforming growth factor-beta 1 gene. *J Biol Chem*. 1989 Jan 5;264(1):402-8.
29. Yamada Y, Miyauchi A, Takagi Y, Tanaka M, Mizuno M, Harada A. Association of the C-509-->T polymorphism, alone or in combination with the T869-->C polymorphism, of the transforming growth factor-beta1 gene with bone mineral density and genetic susceptibility to osteoporosis in Japanese women. *J Mol Med*. 2001 Apr;79(2-3):149-56.
30. Grainger DJ, Heathcote K, Chiano M, Snieder H, Kemp PR, Metcalfe JC, et al. Genetic control of the circulating concentration of transforming growth factor type beta1. *Hum Mol Genet*. 1999 Jan;8(1):93-7.

Reprint Permissions

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Apr 29, 2010

This is a License Agreement between Jasmine Healy ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2418261414055
License date	Apr 29, 2010
Licensed content publisher	John Wiley and Sons
Licensed content publication	International Journal of Cancer
Licensed content title	Functional impact of sequence variation in the promoter region of TGFB1
Licensed content author	Healy Jasmine, Dionne Jo
Licensed content date	Apr 23, 2009
Start page	1483
End page	1489
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Order reference number	
Total	0.00 USD

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a "Wiley Company") or a society for whom a Wiley Company has exclusive publishing rights in relation to a particular journal (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your Rightslink account (these are available at any time at <http://myaccount.copyright.com>).

Terms and Conditions

1. The materials you have requested permission to reproduce (the "Materials") are protected by copyright.

2. You are hereby granted a personal, non-exclusive, non-sublicensable, non-transferable, worldwide, limited license to reproduce the Materials for the purpose specified in the licensing process. This license is for a one-time use only with a maximum distribution equal to the number that you identified in the licensing process. Any form of republication granted by this licence must be completed within two years of the date of the grant of this licence (although copies prepared before may be distributed thereafter). Any electronic posting of the Materials is limited to one year from the date permission is granted and is on the condition that a link is placed to the journal homepage on Wiley's online journals publication platform at www.interscience.wiley.com. The Materials shall not be used in any other manner or for any other purpose. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher and on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Material. Any third party material is expressly excluded from this permission.

3. With respect to the Materials, all rights are reserved. No part of the Materials may be copied, modified, adapted, translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Materials without the prior permission of the respective copyright owner. You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Materials, or any of the rights granted to you hereunder to any other person.

4. The Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc or one of its related companies (WILEY) or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto.

5. WILEY DOES NOT MAKE ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND WAIVED BY YOU.

6. WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

7. You shall indemnify, defend and hold harmless WILEY, its directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

8. IN NO EVENT SHALL WILEY BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

9. Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and

enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

10. The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

11. This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

12. These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in a writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

13. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

14. WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

15. This Agreement shall be governed by and construed in accordance with the laws of England and you agree to submit to the exclusive jurisdiction of the English courts.

BY CLICKING ON THE "I ACCEPT" BUTTON, YOU ACKNOWLEDGE THAT YOU HAVE READ AND FULLY UNDERSTAND EACH OF THE SECTIONS OF AND PROVISIONS SET FORTH IN THIS AGREEMENT AND THAT YOU ARE IN AGREEMENT WITH AND ARE WILLING TO ACCEPT ALL OF YOUR OBLIGATIONS AS SET FORTH IN THIS AGREEMENT.

V1.2

Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK10776283.

Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

**Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006**

If you find copyrighted material related to this license will not be used and wish to cancel, please contact us referencing this license number 2418261414055 and noting the reason for cancellation.

Replication Analysis Confirms the Association of *ARID5B* with Childhood B-Cell Acute Lymphoblastic Leukemia

Jasmine Healy,¹ Chantal Richer,¹ Mathieu Bourgey,¹ Ekaterini A. Kritikou,¹
and Daniel Sinnett^{1,2}

¹Division of Hematology-Oncology, Sainte-Justine Hospital Research Center,
University of Montreal, QC, Canada;

²Department of Pediatrics, Faculty of Medicine, University of Montreal, QC,
Canada

This research was originally published in *Haematologica the Hematology Journal*. Healy J, Richer C, Bourgey M, Kritikou EA, Sinnett D. Replication analysis confirms the association of *ARID5B* with childhood B-cell acute lymphoblastic leukemia. *Haematologica*. 2010; Epub ahead of print, doi:10.3324/haematol.2010.022459. © the Ferrata Storti Foundation. The article is included with the permission of the authors and the journal.

Author Contributions

For this article, I carried out 80% of the work. I participated in the conception and design of the study and performed all the statistical analyses. I also wrote the manuscript. C. Richer, E. Kritikou and D. Sinnett also contributed to the design of the study. C. Richer performed the genotyping. M. Bourgey contributed to the interpretation of the data. E. Kritikou was involved in critical manuscript revision and editing. D. Sinnett is the principal investigator and participated in the coordination of the study as well as writing the manuscript. All authors approved the final version of the manuscript.

Abstract

Although childhood acute lymphoblastic leukemia (ALL) is the most common pediatric cancer, its etiology remains poorly understood. In an attempt to replicate the findings of two recent genome-wide association studies in a French-Canadian cohort, we confirmed the association of five SNPs (rs7073837 ($p = 4.2 \times 10^{-4}$), rs10994982 ($p = 3.8 \times 10^{-4}$), rs10740055 ($p = 1.6 \times 10^{-5}$), rs10821936 ($p = 1.7 \times 10^{-7}$) and rs7089424 ($p = 3.6 \times 10^{-7}$)) in the *ARID5B* gene with childhood ALL. We also confirmed a selective effect for B-cell ALL with hyperdiploidy and report a putative gender-specific effect of *ARID5B* SNPs on ALL risk in males. This study provides a strong rationale for more detailed analysis to identify the causal variants at this locus and to better understand the overall functional contribution of *ARID5B* to childhood ALL susceptibility.

Introduction

Childhood acute lymphoblastic leukemia (ALL), the leading cause of cancer-related deaths among children, is a heterogeneous disease with subtypes that differ markedly in their cellular and molecular characteristics. Advances in our understanding of the pathobiology of ALL have led to risk-targeted treatment regimes and increased long-term survival rates (1). Yet, approximately 20% of patients do not respond to current treatment protocols, and over two-thirds of the survivors experience long-term treatment-related health problems (2).

The etiology of pediatric ALL remains poorly understood. Initiation of leukemogenesis occurs during fetal life or in early infancy and is likely caused by multiple environmental and genetic factors (3). The assertion that ALL may have a genetic basis has long been pursued through association studies based on candidate genes; genes involved in xenobiotic metabolism (4), oxidative stress response (5), DNA repair (6), folate metabolism (7) and cell cycle regulation (8) have been associated with ALL. Two recent genome-wide association studies (GWAS) have provided convincing evidence that inherited genetic variation contributes to childhood ALL predisposition (9, 10). Using different genotyping platforms, Illumina Infinium HD Human 370Duo BeadChips (9) and Affymetrix 500K Arrays (10), both studies found strong associations between variants at 10q21.2 (*ARID5B*) and 7p12.2 (*IKZF1*) and childhood ALL risk. Both *ARID5B* and *IKZF1* are involved in transcriptional regulation and differentiation of B-lymphocyte progenitors. These studies also pointed to *CEBPE* (14q11.2), *DDC* (7p12.2) and *OR2C3* (1q44) as potential ALL susceptibility loci and indicated that common germline variants within the five loci identified may be associated with specific ALL subtypes. Follow-up studies confirmed that variants at loci 10q21.2, 7p12.2 and 14q11.2 are involved in B-cell ALL (11). Variation in *ARID5B* was shown to contribute to ALL risk across

different racial groups (12) further highlighting the importance of this gene in the etiology of childhood ALL.

We attempted to replicate 15 of the initial GWAS signals from the Papaemmanuil et al. and Trevino et al. studies in a French-Canadian cohort consisting of 284 B-precursor ALL cases and 270 healthy controls from the Quebec Childhood ALL (QcALL) study. We replicated the association of 5 SNPs within the *ARID5B* gene, further confirming the implication of this gene in B-cell ALL. Our work provides strong rationale for additional studies to identify the causal variants at this candidate risk locus. This is the first replication study that has attempted to replicate association signals from both initial GWASs in an independent population and the first to report a putative gender-specific effect of *ARID5B* on ALL risk in males.

Design and Methods

Study subjects

Our cohort consisted of 284 childhood B-cell ALL patients and 270 healthy controls. In addition, parental DNA was available for 203 of the probands. Study subjects were French-Canadians of European descent from the established Quebec Childhood ALL (QcALL) cohort (see (4, 8)). Briefly, incident childhood pre-B ALL cases were diagnosed in the Hematology-Oncology Unit of Sainte-Justine Hospital, Montreal, Canada, between October 1985 and November 2006. The current study sample includes 170 males and 114 females with a median age of 4.2 years. This patient cohort is representative of the childhood pre-B ALL population; for clinical characteristics of cases, see Table 1. Healthy controls, 152 males and 118 females with a median age of 30.1 years,

consisted of a group of newborns and adults recruited through clinical departments other than the Hematology-Oncology Unit, Sainte-Justine Hospital. Peripheral blood or bone marrow (samples in remission) was collected from all participants and DNA was extracted as previously described (13). The Institutional Review Board approved the research protocol and informed consent was obtained from all participants and/or their parents.

SNP genotyping and quality control checks

SNPs were genotyped using the Luminex xMAP/Autoplex Analyser CS1000 system (Perkin Elmer, Waltham, MA). The fifteen selected SNPs were amplified in a single multiplexed assay and hybridized to Luminex MicroPlex™ –xTAG Microspheres (14) for genotyping using allele-specific primer extension (ASPE). The PCR and TAG-ASPE primers are shown in Supplementary Table 1; amplification and reaction conditions are available upon request. Allele calls were assessed and compiled using the Automatic Luminex Genotyping software (M. Bourgey et al., manuscript submitted, 2009). The average genotype call rate for the 15 SNPs was 97.0%. Hardy-Weinberg equilibrium (HWE) was tested using the X^2 goodness of fit test and PedCheck (Version 1.1) was used to identify genotype incompatibilities using the familial data (15); inconsistent case-parent trios were removed from the analysis.

Statistical analysis

Statistical analyses were done using STATA/IC Version 10.1 (StataCorp, College Station, TX). Pearson's X^2 test or Fisher's exact test, as appropriate, was used to compare allele/genotype/haplotype carriership in patients and controls. Crude odds ratios (ORs) were measured using logistic regression and are given with 95% confidence intervals (CIs). Pairwise linkage disequilibrium (LD) estimates were measured in STATA. We assessed gender-specific

associations through stratified analysis comparing male cases to male controls or female cases to female controls and the Mantel-Haenszel (MH) chi-square test of homogeneity was used to test for significant risk differences between males and females. Haplotype reconstruction was performed using the FAMHAP Software (Version 16), using parental data when available (16); incorporating genotype information of related individuals increases precision of haplotype reconstruction and frequency estimation (16-18). Logistic regression was used to estimate haplotype-specific ORs using the most common haplotype as reference and a likelihood ratio test implemented in FAMHAP was used to test for global haplotype association with disease status. Multiple testing corrections were performed using the Benjamini-Hochberg false discovery rate (FDR) method with a type I error rate of 5%; nominal p values are shown.

Results and Discussion

We genotyped the top ten SNPs from Papaemmanuil et al. (GWA1) (9) and five SNPs from Trevino et al. (GWA2) (10) in a French-Canadian cohort of European descent. The distribution of genotype frequencies in all 15 SNPs were in HWE ($p > 0.05$). Risk allele frequencies were similar to those observed in the European populations of both GWAS (9, 10) (Table 2).

Univariate analysis showed highly significant allelic associations within chromosomal region 10q21.2. The five SNPs from this region annotated the *ARID5B* gene and were strongly associated with B-cell ALL risk in our cohort; OR estimates were in the same direction and were of similar strength as those previously reported (Table 2). SNPs rs10994982, rs10740055, rs10821936 and rs7089424 span a 42kb region in intron 3 of *ARID5B* whereas SNP rs7073837 is located in intron 2. rs10821936, the strongest association signal from GWA2, was the most significant signal in our study ($p = 1.7 \times 10^{-7}$). rs10821936 ($p =$

3.6×10^{-7}) is in strong LD with rs7089424 ($r^2 = 0.95$), the second-strongest association signal in our study. SNPs rs10994982 (GWA1) and rs10740055 (GWA2) were highly correlated ($r^2 = 0.93$) and strongly associated with childhood B-cell ALL ($p = 3.8 \times 10^{-4}$ and $p = 1.6 \times 10^{-5}$, respectively). rs7073837 was in moderate LD with SNP pairs rs10821936-rs7089424 and rs10994982-rs10740055 (r^2 of 0.65 and 0.72, respectively) and was also associated with disease ($p = 4.2 \times 10^{-4}$). These 5 SNP associations withstood multiple testing corrections and remained significant after controlling for a false discovery rate of 5%. Using subtype analysis, we confirmed that *ARID5B* SNPs were significantly associated with B-hyperdiploid ALL (p values $\leq 2.0 \times 10^{-4}$) (Table 2 and (9)).

The direction and effect sizes were replicated for most of the SNPs tested; out of the 15 SNPs interrogated in this study only one (*OR2C3* rs1881797) had an OR which tended in the opposite direction (93%; $P = 0.00049$; binomial probability). However the association signals at the *IKZF1*, *DDC*, *CEBPE*, and *OR2C3* loci failed to reach statistical significance in our cohort (Table 2). Lack of confirmation of association with chromosomal region 7p12.2 was surprising given the strong statistical association observed in both original GWASs, and the convincing support of a recent follow-up study conducted in a large German case-control cohort for *IKZF1* (11). Risk allele frequencies in cases did not differ between cohorts therefore failure to replicate is unlikely due to genetic heterogeneity. The most likely explanation for the lack of replication is the limited power of our study to detect loci with weaker effects. With our limited sample size, we had 80% power at the 5% level, to detect a minimum OR of 1.8 with RAFs $\geq 20\%$ and of 2.1 with RAFs $\geq 10\%$. Lack of replication could also partially reflect the complexity underlying ALL pathogenesis; for example, molecular characterization of the disease might differ across studies. Further association studies with larger case-control samples and detailed subgroup analysis are required to investigate whether the associations between 7p12.2 (*IKZF1* and *DDC*) and 14q11.2 (*CEBPE*) hold true.

To further describe the observed *ARID5B* associations and capture associations under various genetic models, we measured the corresponding genotype ORs in all samples, as well as in males and females separately. Carriers of a homozygous risk genotype at SNPs rs7073837, rs10994982 and rs10740055 had over a 2-fold increase in B-cell ALL risk. A strong allele dose-dependent effect on risk was observed at loci rs10821936 and rs7089424 ($p_{\text{trend}} = 7.4 \times 10^{-7}$ and 1.7×10^{-6} , respectively) (Table 3). Significant risk differences were found between males and females at loci rs10994982 and rs10740055: a 3.8-fold and 4.4-fold increase in risk was observed in male carriers of the homozygous risk genotypes, respectively, while no significant effect was observed in females (MH p values < 0.03) (Table 3). Although the effects of rs10821936 and rs7089424 were more marked in males, the gender difference was not significant at these loci (p values ≥ 0.15).

We observed similar gender-biases among hyperdiploid ALL patients for variants rs10994982 (Males AA vs. GG: OR(95%C.I.)= 6.25(2.40-17.49); Females AA vs. GG: OR(95%C.I.)= 1.12(0.33-3.96); MH $p= 0.017$), and rs10740055 (Males CC vs. AA: OR(95%C.I.)= 6.90(2.43-22.18); Females CC vs. AA: OR(95%C.I.)= 1.66(0.49-6.14); MH $p= 0.059$) (data not shown). However, the wide confidence intervals caused by overstratification of the data emphasize the uncertainty of the risk estimates.

Finally, we performed multivariate haplotype analysis for *ARID5B* (Supplementary Table 2). 15 different haplotypes could be inferred, but only four haplotypes had frequencies ≥ 0.05 and represented ~96% of the observed haplotypes in our sample. The remaining 4% of the chromosomes carried eleven minor haplotypes. We found a significant difference in the overall distribution of the 15 *ARID5B*-derived haplotypes between B-cell ALL cases and controls (Global $\chi^2 = 45.03$, 14 degrees of freedom, $p = 4.0 \times 10^{-5}$). The most

frequent haplotype among controls (50%) carried non-risk alleles at all five of the *ARID5B* loci (CGATT) whereas the complementary haplotype, formed by the risk alleles of these polymorphisms (AACCG), was the most abundant haplotype among cases (46.25%). Through haplotype-specific tests we showed that the risk haplotype AACCG was associated with a near 2-fold increase in B-cell ALL susceptibility ($OR(95\%CI) = 1.93(1.47-2.53)$, $p = 7.6 \times 10^{-7}$) (Supplementary Table 2). And when stratified by gender, similar results were observed in the male subgroup only. The haplotype analysis further demonstrates that the associations observed at the 5 individual *ARID5B* loci are not independent and likely reflect a single association signal.

rs10821936 was the strongest association signal in our study ($p = 1.7 \times 10^{-7}$). To further verify that the associations detected at *ARID5B* are not independent, we performed logistic regression analysis on each of the additional SNPs, adjusting for the effect of variant rs10821936 (data not shown). No residual association was provided by the remaining four SNPs after accounting for the effect of rs10821936 and the likelihood ratio tests, comparing the reduced univariate rs10821936 model to each of the extended 2-SNP models, were all non-significant ($p > 0.12$), suggesting no evidence of multiple effects operating across the *ARID5B* SNPs.

Our replication data confirms that *ARID5B* is a novel susceptibility factor for childhood B-cell ALL and corroborate previous findings of a putative selective effect for B-cell precursor ALL with hyperdiploidy. We also report a gender-specific effect of *ARID5B* SNPs on ALL risk in males. *ARID5B* plays a vital role in the regulation of embryonic development and cell growth and differentiation through tissue-specific repression of differentiation-specific gene expression (19, 20). Aberrant *ARID5B* expression in the developing fetus could halt B-lymphocyte maturation and contribute to leukemogenesis. B-cell ALL incidence

is higher in males and though our data suggest a gender bias in the effect of *ARID5B* variation on disease risk, the link between *ARID5B* and increased risk of leukemia among males remains to be determined. Given the combined statistical significance of association of this region, re-sequencing and functional analyses are now required to identify the causal variants at the 10q21.2 locus. Better elucidation of the mechanisms through which *ARID5B* variants are involved in childhood ALL could be of great diagnostic value and help guide risk-directed therapy, ultimately improving disease management and outcome.

Tables

Table 1. Characteristics of the B-cell ALL patients from the Quebec Childhood ALL cohort

Patient Characteristics	Cases, n (%)
Total number of subjects	284
Gender	
Male	170 (59.9)
Female	114 (40.1)
Age group, years	
≤ 1	6 (2.1)
1-10	201 (70.8)
> 10	41 (14.4)
N/D	36 (12.7)
Hyperdiploid	
Positive	106 (37.3)
Negative	156 (54.9)
N/D	22 (7.7)
Chromosomal translocations	
Absence of translocation	92 (32.4)
t(12;21)	35 (12.3)
Other	12 (4.2)
N/D	145 (51.1)

N/D indicates no data available

Table 2. Replication analysis in the Quebec Childhood ALL cohort of germline SNPs whose allele frequencies differed between children with ALL and control groups in two genome-wide association studies

Gene (Chr.), DNA variant	Risk allele	QcALL replication study				GWA1 - Papaemmanuil et al.				GWA2 - Trevino et al.			
		RAF	Subgroup	OR (95% CI)	<i>p</i> value	RAF	Subgroup	OR (95% CI)	<i>p</i> value	RAF	Subgroup	OR (95% CI)	<i>p</i> value
ARID5B (10q21.2)													
rs7073837	A	0.34	B-cell	1.54 (1.21-1.97)	4.2E-04	0.40	—	1.58 (1.35-1.89)	4.7E-16			—	
			B-hyperdip	2.10 (1.51-2.93)	9.6E-06		B-cell	1.59 (1.48-1.71)	1.0E-15				
rs10994982	A	0.48	B-cell	1.55 (1.22-1.97)	3.8E-04			—		0.47	—	1.71 (1.43-2.05)	1.2E-09
			B-hyperdip	1.87 (1.34-2.60)	2.0E-04						B-hyperdip	1.71 (1.19-2.46)	0.0025
rs10740055	C	0.49	B-cell	1.71 (1.34-2.19)	1.6E-05	0.50	—	1.53 (1.41-1.64)	5.4E-14			—	
			B-hyperdip	2.06 (1.46-2.90)	3.0E-05		B-cell	1.57 (1.45-1.81)	1.6E-14				
rs10821936	C	0.33	B-cell	1.93 (1.51-2.48)	1.7E-07			—		0.33	—	2.00 (1.68-2.38)	2.8E-16 ^a
			B-hyperdip	2.91 (2.08-4.07)	2.0E-10						B-hyperdip	2.12 (1.49-3.01)	9.7E-06
rs7089424	G	0.43	B-cell	1.91 (1.49-2.45)	3.6E-07	0.34	—	1.65 (1.54-1.76)	6.7E-19			—	
			B-hyperdip	2.87 (2.04-4.04)	8.2E-10		B-cell	1.70 (1.58-1.81)	1.4E-19				
							B-hyperdip	—	3.8E-06 ^b				
IKZF1 (7p12.2)													
rs6964823	G	0.54	B-cell	1.15 (.90-1.46)	0.27	0.50	—	1.52 (1.41-1.64)	6.0E-14			—	
							B-cell	1.53 (1.42-1.65)	1.9E-13				
rs11978267	G	0.31	B-cell	1.27 (.99-1.65)	0.065			—		0.27	—	1.69 (1.4-1.9)	8.8E-11
rs4132601	G	0.31	B-cell	1.26 (.98-1.62)	0.075	0.28	—	1.69 (1.58-1.81)	1.2E-19 ^c			—	
							B-cell	1.73 (1.61-1.85)	9.3E-20				
rs6944602	A	0.24	B-cell	1.17 (.89-1.55)	0.27	0.21	—	1.64 (1.37-2.07)	3.4E-15			—	
							B-cell	1.69 (1.56-1.81)	1.5E-15				

DDC (7p12.2)											
rs7809758	G	0.41	B-cell	1.02 (.80-1.30)	0.9	0.37	—	1.44 (1.32-1.54)	2.4E-10		
							B-cell	1.48 (1.37-1.60)	2.9E-11		—
rs880028	C	0.25	B-cell	1.04 (.79-1.36)	0.79	0.22	—	1.43 (1.30-1.56)	1.3E-07		—
							B-cell	1.49 (1.36-1.61)	1.4E-09		
rs3779084	C	0.25	B-cell	1.01 (.77-1.33)	0.92	0.22	—	1.44 (1.32-1.56)	8.8E-09		—
							B-cell	1.50 (1.37-1.63)	6.5E-10		
rs2242041	G	0.11	B-cell	1.40 (.99-2.00)	0.059			—		0.09	—
CEBPE (14q11.2)											
rs2239633	G	0.53	B-cell	1.19 (.94-1.52)	0.15	0.52	—	1.34 (1.22-1.45)	2.9E-07		—
							B-cell	1.37 (1.26-1.49)	5.6E-08		
OR2C3 (1q44)											
rs1881797	C	0.17	B-cell	0.95 (.66-1.35)	0.77			—		0.16	—
			t(12;21)	1.31 (.53-3.24)	0.57					t(12;21)	1.52 (1.2-1.8)
											7.3E-06
											2.08 (1.1-3.8)
											0.021

Results are shown for allelic case-control association tests; *p* values in bold remain significant after FDR adjustment for multiple testing at the 5% level. The Quebec Childhood ALL (QcALL) replication cohort consisted of 284 B-cell ALL cases and 270 controls; logistic regression was applied to either the full dataset or to a restricted subgroup of patients as specified. GWA1; the Papaemmanuil et al. study comprised 907 cases (824 B-cell, 83 T-cell) and 2,398 controls; logistic regression was applied to either the full dataset or to a restricted subgroup of patients as specified; *P* values denote Cochran-Armitage trend test statistics. GWA2; the Trevino et al. study consisted of a discovery cohort of 317 cases (274 B-cell, 43 T-cell) and 17,958 controls; logistic regression was used and subgroup analysis was performed by comparing allele frequencies between single ALL subgroups and all other subgroups combined. RAF indicates risk allele frequency in controls; OR, odds ratio; CI, confidence interval; —, not applicable.

^a Strongest association signal from the Trevino et al. study.

^b *p* value denotes case-only logistic regression analysis.

^c Strongest association signal from the Papaemmanuil et al. study.

Table 3. Distribution of *ARID5B* genotypes among B-cell ALL cases and controls from the Quebec Childhood ALL cohort and gender-specific genotype risks estimates

DNA variant, Genotype	No. (%)		OR (95% CI)	p value	Males/Females (%)		OR (95% CI)		p value ^a
	B-cell ALL Cases	Controls			B-cell ALL Cases	Controls	Males	Females	
rs7073837									
CC	67 (24.81)	93 (35.23)	1 (referent)	—	39 (23.8)/28 (26.4)	59 (39.9)/34 (29.6)	1 (referent)	1 (referent)	—
CA	128 (47.41)	127 (48.11)	1.40 (0.94-2.08)	0.1	75 (45.7)/53 (50.0)	65 (43.9)/61 (53.0)	1.75 (1.00-3.05)	1.05 (.54-2.06)	0.22
AA	75 (27.78)	44 (16.67)	2.37 (1.45-3.85)	4.8E-04	50 (30.5)/25 (23.6)	24 (16.2)/20 (17.4)	3.15 (1.60-6.25)	1.52 (.65-3.53)	0.15
rs10994982									
GG	50 (18.18)	72 (27.27)	1 (referent)	—	28 (16.8)/22 (20.4)	50 (33.1)/22 (19.6)	1 (referent)	1 (referent)	—
GA	125 (45.45)	129 (48.86)	1.39 (0.90-2.16)	0.14	71 (42.5)/54 (50.0)	69 (45.7)/59 (52.7)	1.84 (1.00-3.39)	.92 (.43-1.95)	0.13
AA	100 (36.36)	63 (23.86)	2.29 (1.42-3.69)	6.6E-04	68 (40.7)/32 (29.6)	32 (21.2)/31 (27.7)	3.79 (1.94-7.45)	1.03 (.45-2.39)	0.01
rs10740055									
AA	41 (15.41)	67 (25.48)	1 (referent)	—	22 (13.7)/19 (18.1)	44 (29.3)/23 (20.5)	1 (referent)	1 (referent)	—
AC	117 (43.98)	132 (50.19)	1.45 (0.91-2.30)	0.12	65 (40.4)/52 (49.5)	72 (48.0)/59 (52.7)	1.81 (.94-3.51)	1.07 (.49-2.32)	0.27
CC	108 (40.60)	64 (24.33)	2.76 (1.68-4.53)	5.0E-05	74 (46.0)/34 (32.4)	34 (22.7)/30 (26.8)	4.35 (2.16-8.84)	1.37 (.58-3.23)	0.03
rs10821936									
TT	76 (27.94)	127 (48.47)	1 (referent)	—	42 (25.8)/34 (31.2)	72 (48.3)/55 (49.1)	1 (referent)	1 (referent)	—
TC	129 (47.43)	99 (37.79)	2.18 (1.48-3.20)	7.2E-05	73 (44.8)/56 (51.4)	56 (37.6)/42 (37.5)	2.23 (1.29-3.87)	2.16 (1.15-4.04)	0.93
CC	67 (24.63)	36 (13.74)	3.11 (1.90-5.10)	4.8E-06	48 (29.4)/19 (17.4)	21 (14.1)/15 (13.4)	3.92 (1.98-7.84)	2.05 (.85-4.95)	0.21
rs7089424									
TT	62 (23.85)	115 (45.28)	1 (referent)	—	34 (21.7)/28 (27.2)	64 (44.1)/51 (47.2)	1 (referent)	1 (referent)	—
TC	131 (50.38)	99 (38.98)	2.45 (1.63-3.68)	1.1E-05	75 (47.8)/56 (54.4)	59 (40.7)/39 (36.1)	2.39 (1.35-4.25)	2.62 (1.35-5.08)	0.83
CC	67 (25.77)	40 (15.75)	3.11 (1.89-5.12)	6.1E-06	48 (30.6)/19 (18.4)	22 (15.2)/18 (16.7)	4.11 (2.04-8.35)	1.92 (.81-4.58)	0.15

Logistic regression was used to estimate ORs in either the full dataset or in restricted subgroups stratified by gender, comparing male cases to male controls or female cases to female controls. Percentages indicate number of individuals with a given genotype/total number of genotyped individuals. *p* values in bold remain significant after FDR adjustment for multiple testing at the 5% level. OR indicates odds ratio; CI, confidence interval; —, not applicable.

^a *p* value denotes the Mantel-Haenszel chi-square test of homogeneity for risk differences between males and females.

Supplementary Table 1. Summary of primers used in the PCR and allele-specific primer extension (ASPE) assays for SNP genotyping

Gene (Chr)	DNA variant	Position	PCR primers	Product size (bp)	Allele	TAG-ASPE primers
ARID5B (10q21.2)						
	rs7073837	63,369,901	F: ACCAGAATGCACACAGTCTCCTTGC R: GGTCCAGCGTGGAAGCCACA	241	A C	TCATTTACCTTTAATCCAATAATCTGGAGAGTGGATCATTCCCTCA TCAATTACCTTTTCAATAACAATACGGAGAGTGGATCATTCCCTCC
	rs10994982	63,380,110	F: ACCTCGTGATCTGCCCCGCT R: CCACCTCGGCTTCCGGAGT	204	A G	CTAATTACTAACATCACTAACAAATCATGGTCTTTTAAATATCTTTTGAGAATGCAA TCAATTACTTCACTTTAATCCTTTTCATGGTCTTTTAAATATCTTTTGAGAATGCAG
	rs10740055	63,388,485	F: ACCACTATGCACCTTATCGGAGACAACA R: GCCCGGCCGTGACCTCTTTT	278	G T	CAATAAACTATACTTCTTCACTAACACAGGGTTTCTATTGGAAAGCTGG CTTTAATCCTTTATCACTTTATCACACAGGGTTTCTATTGGAAAGCTGT
	rs10821936	63,393,583	F: ACCAGCTCGGCAGAGCATCC R: CCCCGGTGCTTGAACACACT	257	C T	TCATAATCTCAACAATCTTTCTTTCTCTGTGTGCAGTTACTATAGTTGTAC TCATCAATCAATCTTTTCACTTTCTCTGTGTGCAGTTACTATAGTTGTAT
	rs7089424	63,422,165	F: TGGCTCCCGGTCTGTGGCTTA R: CCCCCAACCAAGGGTTTGAACCAGC	271	G T	TACACTTCTTTCTTTCTTTCTTTGGTAGCAGTGTGGTTATAGTTTAGTTG CAATTTATCATTCATTCAATTTTCCAGGTAGCAGTGTGGTTATAGTTAGTTT
IKZF1 (7p12.2)						
	rs6964823	50,234,305	F: CTGTCCCTGCAGCGTCTGGC R: AGGCCTGGCCGGCTTCTCAG	289	C T	TCATCAATCTTTCAAATTTACTTTACGACGAATGGTCTTGTCTTCTTCTTC TTACTTCACTTTCTATTTACAATCAGACGAATGGTCTTGTCTTCTTCTTT
	rs11978267	50,240,513	F: ACTGGCAAGACTGCGGCTGTG R: GCTCTCATGGCACGCTCCCC	298	C T	CAATTCAAAATCAAAATCAATCAGGGGAGGGTAGGTAGAAGTTTATGC AATCCTTTCTTTAATCTCAAACTACTAGGGGAGGGTAGGTAGAAGTTTATGT
	rs4132601	50,244,813	F: GGGTGTGGCATTGGAACGGGA R: ACCAACTGACTCAGGGGATGGA	231	G T	TCAAAATCTCAAACTACTCAAAATCATCACAGAGAAAGATGCGCCTG CAATTAACATACATAAATACATACATCACAGAGAAAGATGCGCCTT
	rs6944602	50,247,960	F: ACACAGCTTCACGGTGTGACCC R: ATTGCTGCACGGGGTCTGC	184	A G	TACACAACTTTTTCATTACATCATCCAGCGGGCATTACACTCA ATCATACATACATACAAATCTACACCAGCGGGCATTACACTCG
DDC (7p12.2)						
	rs7809758	50,347,542	F: TGTCAGCCAGCCACCTGT R: CCCAGTTCCGAGACCTTTGTGGC	180	C T	CTTTATCAATACATACTACAATCAAAAGCTGGTGAGACGAACCTTCC ATACTACATCATAATCAAAATCAAAAGCTGGTGAGACGAACCTTCT
	rs880028	50,344,345	F: ACTTGGCCAGGGGACAGCA R: TCCTGGCCAGTAGCTGGCTGAT	106	A G	AATCTTACTACAAATCCTTTCTTTGCCCTTTAAATCCGATAGCCCTA CTTTTCATCAATAATCTTACCTTTGCCCTTTAAATCCGATAGCCCTG
	rs3779084	50,342,944	F: GGGATGCATGGAGCTGTGGGC R: CCCCACTGAGGCAGCCTTGC	190	A G	CTTTTCAAAATCAAACTACTCAACTTTGCCCATCTAGGAACCAAGGCA CTACAAACAAACAAACATTATCAAGCCCATCTAGGAACCAAGGCG
	rs2242041	50,303,658	F: ACAGCGGTACTTTCCCTCCCT R: CTCAAGCATCTCTGGAAGGTAGTGGGG	99	C G	CTATCTTTAAACTACAAATCTAACGCTTATGCTGAGAGCAATGGAATAAC CTAACTAACAAATCTAACTAACGCTTATGCTGAGAGCAATGGAATAAG
CEBPE (14q11.2)						
	rs2239633	22,658,897	F: TGCTGGGCTCCACCTACCCC R: CTCTGGAGCACCCACGCAGGC	159	A G	TCATTTACAAATTCAAATTAAGTCTAGGAACAAGCTCTACACA CAATTCATTTACCAATTTACCAATTAAGTCTAGGAACAAGCTCTACACG
OR2C3 (1q44)						
	rs1881797	244,015,573	TGTGCACAGCATGTGTTGAGATGA GCACTGGACACAGCTTCTGCCT	292	C T	CTACTATACATCTTACTATACTTTTGGGTTAGTTTCTGAATTTCTGCATT TTACCTTTTATACCTTTCTTTTACTGGGTTAGTTTCTGAATTTCTGCATT

^a SNP position relative to the UCSC Genome Browser Human May 2004 Assembly (hg17)

Supplementary Table 2. Distribution of *ARID5B* haplotypes among B-cell ALL cases and controls from the Quebec Childhood ALL cohort and gender-specific haplotype risks estimates

Haplotype	DNA variant					B-cell ALL		OR (95% CI)	p value	Global χ^2 (df)	Global p value
						Cases	Controls				
	rs7073837	rs10994982	rs10740055	rs10821936	rs7089424	Total (%)					
ARID5B*1	C	G	A	T	T	224 (38.23)	268 (50.00)	1 (referent)	—	45.03 (14)	4.04E-05
ARID5B*2	A	A	C	C	G	271 (46.25)	168 (31.34)	1.93 (1.47-2.53)	7.6E-07		
ARID5B*3	A	A	C	T	T	31 (5.29)	46 (8.58)	0.81 (.48-1.35)	0.39		
ARID5B*4	C	A	C	T	T	37 (6.31)	34 (6.34)	1.30 (.77-2.21)	0.30		
ARID5B*	—	—	—	—	—	23 (3.92)	20 (3.73)	1.38 (.70-2.71)	0.31		
Males (%)											
ARID5B*1	C	G	A	T	T	120 (34.29)	159 (52.65)	1 (referent)	—	36.95 (16)	2.13E-03
ARID5B*2	A	A	C	C	G	170 (48.57)	92 (30.46)	2.45 (1.70-3.52)	3.41E-07		
ARID5B*3	A	A	C	T	T	19 (5.43)	21 (6.95)	1.20 (.58-2.46)	0.59		
ARID5B*4	C	A	C	T	T	26 (7.43)	15 (4.97)	2.30 (1.11-4.87)	0.014		
ARID5B*	—	—	—	—	—	15 (4.29)	15 (4.97)	1.32 (.58-3.03)	0.46		
Females (%)											
ARID5B*1	C	G	A	T	T	104 (44.07)	108 (46.35)	1 (referent)	—	21.45 (9)	0.011
ARID5B*2	A	A	C	C	G	99 (41.95)	75 (32.19)	1.37 (.90-2.09)	0.120		
ARID5B*3	A	A	C	T	T	13 (5.51)	24 (10.30)	0.56 (.25-1.22)	0.120		
ARID5B*4	C	A	C	T	T	11 (4.66)	20 (8.58)	0.57 (.23-1.33)	0.16		
ARID5B*	—	—	—	—	—	9 (3.81)	6 (2.58)	1.56 (.47-5.50)	0.41		

Logistic regression was used to estimate haplotype-specific ORs in either the full dataset or in restricted subgroups stratified by gender, comparing male cases to male controls or female cases to female controls. The most common haplotype was used as reference. *p* values in bold remain significant after FDR adjustment for multiple testing at the 5% level. Percentages indicate number of chromosomes with given haplotype/total number of chromosomes. Haplotypes with relative frequencies <5% are grouped under ARID5B* and are represented as * combinations of the four DNA variants.

A likelihood ratio test was performed in FAMHAP to compare global haplotype differences between cases and controls and is reported here as a Global chi-square test with number of haplotype parameters different from zero-1 degrees of freedom. OR indicates crude odds ratio; df, degrees of freedom; and —, not applicable.

^a Significant risk difference between males and females based on Mantel-Haenszel chi-square test of homogeneity: $X^2=7.31$ (1df); p value= 0.0069.

References

1. Pieters R, Carroll WL. Biology and treatment of acute lymphoblastic leukemia. *Pediatr Clin North Am*. 2008 Feb;55(1):1-20, ix.
2. Mody R, Li S, Dover DC, Sallan S, Leisenring W, Oeffinger KC, et al. Twenty-five-year follow-up among survivors of childhood acute lymphoblastic leukemia: a report from the Childhood Cancer Survivor Study. *Blood*. 2008 Jun 15;111(12):5515-23.
3. Greaves M. Molecular genetics, natural history and the demise of childhood leukaemia. *Eur J Cancer*. 1999 Dec;35(14):1941-53.
4. Krajinovic M, Labuda D, Richer C, Karimi S, Sinnett D. Susceptibility to childhood acute lymphoblastic leukemia: influence of CYP1A1, CYP2D6, GSTM1, and GSTT1 genetic polymorphisms. *Blood*. 1999 Mar 1;93(5):1496-501.
5. Krajinovic M, Sinnett H, Richer C, Labuda D, Sinnett D. Role of NQO1, MPO and CYP2E1 genetic polymorphisms in the susceptibility to childhood acute lymphoblastic leukemia. *Int J Cancer*. 2002 Jan 10;97(2):230-6.
6. Batar B, Guven M, Baris S, Celkan T, Yildiz I. DNA repair gene XPD and XRCC1 polymorphisms and the risk of childhood acute lymphoblastic leukemia. *Leuk Res*. 2009 Jun;33(6):759-63.
7. Petra BG, Janez J, Vita D. Gene-gene interactions in the folate metabolic pathway influence the risk for acute lymphoblastic leukemia in children. *Leuk Lymphoma*. 2007 Apr;48(4):786-92.
8. Healy J, Belanger H, Beaulieu P, Lariviere M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood*. 2007 Jan 15;109(2):683-92.
9. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet*. 2009 Sep;41(9):1006-10.
10. Trevino LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet*. 2009 Sep;41(9):1001-5.

11. Prasad RB, Hosking FJ, Vijayakrishnan J, Papaemmanuil E, Kohler R, Greaves MF, et al. Verification of the susceptibility loci on 7p12.2, 10q21.2 and 14q11.2 in precursor B-cell acute lymphoblastic leukemia of childhood. *Blood*. 2009 Dec 30.
12. Yang W, Trevino LR, Yang JJ, Scheet P, Pui CH, Evans WE, et al. ARID5B SNP rs10821936 is associated with risk of childhood acute lymphoblastic leukemia in blacks and contributes to racial differences in leukemia incidence. *Leukemia*. Jan 7.
13. Baccichet A, Qualman SK, Sinnett D. Allelic loss in childhood acute lymphoblastic leukemia. *Leuk Res*. 1997 Sep;21(9):817-23.
14. Koo SH, Ong TC, Chong KT, Lee CG, Chew FT, Lee EJ. Multiplexed genotyping of ABC transporter polymorphisms with the Bioplex suspension array. *Biol Proced Online*. 2007;9:27-42.
15. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet*. 1998 Jul;63(1):259-66.
16. Becker T, Knapp M. Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet Epidemiol*. 2004 Jul;27(1):21-32.
17. Schaid DJ. Relative efficiency of ambiguous vs. directly measured haplotype frequencies. *Genet Epidemiol*. 2002 Nov;23(4):426-43.
18. Schouten MT, Williams CK, Haley CS. The impact of using related individuals for haplotype reconstruction in population studies. *Genetics*. 2005 Nov;171(3):1321-30.
19. Huang TH, Oka T, Asai T, Okada T, Merrills BW, Gertson PN, et al. Repression by a differentiation-specific factor of the human cytomegalovirus enhancer. *Nucleic Acids Res*. 1996 May 1;24(9):1695-701.
20. Wilsker D, Patsialou A, Dallas PB, Moran E. ARID proteins: a diverse family of DNA binding proteins implicated in the control of cell growth, differentiation, and development. *Cell Growth Differ*. 2002 Mar;13(3):95-106.

Reprint Permissions

Yes, sorry for not informing you soon. We grant permission to use the article cited in your message below free of charge.

Best regards

Lorella Ripari

Hello Ms. Ripari,

As a follow-up to our previous emails, would it be possible to please obtain permission to include the HAEMATOL/2010/022459 manuscript in my dissertation. It is now available online for e-pub.

Best regards,

Jasmine Healy

CHAPTER THREE

*COMBINED PATHWAY EFFECTS AND GENE-
GENE INTERACTIONS IN THE
SUSCEPTIBILITY TO CHILDHOOD ACUTE
LYMPHOBLASTIC LEUKEMIA*

In previous analyses, I selected a subset of promoter SNPs (pSNPs) in genes that play a key role in maintaining genomic integrity and cell homeostasis and performed a candidate gene association study to identify single variants and/or haplotypes that may modulate gene expression and thereby influence ALL risk in children. Several of these variants, as well as their associated promoter haplotypes, were shown to be associated with ALL supporting a role for both G1/S cell cycle checkpoint and DNA double-strand break repair mechanisms in disease etiology. However these genes and corresponding pSNPs are not likely to act alone to confer disease susceptibility but rather interact with a number of genes, and possibly environmental factors, and operate through complex biological mechanisms to increase susceptibility. Although we cannot be certain that the identified pSNPs play a functional role in the biological process in which the corresponding gene products are implicated, our genetic data and functional studies suggest that they are likely to affect the overall outcome of the pathways in which they lie. In which case, each individual variant may only have a small incremental effect on disease risk but the genetic effect of combinations of functionally relevant SNPs among several pathway components could lead to dysregulation of the corresponding pathway and contribute to modified childhood ALL risk. The effects of gene-gene or epistatic interactions, whereby the effect of a gene can be altered by one or several other genes, could also dictate disease outcome over the independent effects of any one susceptibility gene.

In the second aim of my study, I used a comprehensive pathway-driven approach to identify genes and pathways that are strongly associated with childhood ALL. In an attempt to better explain the underlying complexities of disease etiology, I also investigated putative gene-gene interaction effects. I applied a semi-Bayesian method which involved the use of hierarchical models to integrate relevant biological and functional information directly into the analytical framework to inform association testing. This is the first time this

approach has been used to successfully identify biological pathways involved in childhood ALL susceptibility, shedding light on the genetic underpinnings of this disease. This work has recently been submitted for publication and is presented in this chapter:

- Promoter variants in genes involved in the cell cycle and DNA repair pathways and the susceptibility to childhood acute lymphoblastic leukemia (Healy et al. (2010) *Cancer Research*, submitted);

Promoter Variants in Genes Involved in the Cell Cycle and DNA Repair Pathways and the Susceptibility to Childhood Acute Lymphoblastic Leukemia

Jasmine Healy,¹ Chantal Richer,¹ Joëlle Dionne,¹ Hélène Bélanger,¹ Mathieu Larivière,¹ Manon Ouimet,¹ Vincent Gagné,¹ Verena Weth,¹ Khadija Benhamza,¹ Patrick Beaulieu,¹ Hugues Massé,¹ Mathieu Bourgey,¹ Marie-Helene Roy-Gagnon,^{1,2} Daniel Sinnett^{1,3}

¹Division of Hematology-Oncology, Sainte-Justine Hospital Research Center, University of Montreal, QC, Canada;

²Department of Social and Preventive Medicine, Faculty of Medicine, University of Montreal, QC, Canada;

³Department of Pediatrics, Faculty of Medicine, University of Montreal, QC, Canada

This research has been submitted to *Cancer Research* and is pending review. The manuscript is included with the permission of the authors.

Author Contributions

For this article, I carried out 80% of the work. I participated in the design of the project and performed some of the genotyping. Using already available methods and software, I performed all the statistical analyses. I also wrote the manuscript. M.H. Roy-Gagnon and D. Sinnett contributed to the conception and design of the study. C. Richer and K. Benhamza performed some of the genotyping. J. Dionne, M. Larivière, M. Ouimet, V. Gagné and V. Weth performed the functional assays while P. Beaulieu was involved in the *in silico* analysis. M. Bourgey and H. Massé contributed to the interpretation of the data. D. Sinnett is principal investigator. All authors approved the final version of the manuscript.

Abstract

Childhood acute lymphoblastic leukemia (ALL) is likely caused by multiple genetic and environmental factors however the identification of established genetic risk factors for ALL has been impeded by its complex and heterogeneous nature. In this study, we postulated that a combination of *cis*-acting sequence variants in multiple genes sharing functions in the G1/S cell cycle checkpoint and DNA double-strand break repair pathways could influence interindividual variability in the susceptibility to childhood ALL by modulating gene expression and affecting the overall outcome of these core regulatory processes. Hierarchical modeling was used to investigate gene- and pathway-based associations between 46 promoter SNPs in 12 cell cycle genes and 7 DNA repair genes with childhood ALL and to explore putative gene-gene interactions, while integrating prior biological and functional information into the analysis. We found that variants in genes *CDKN2A* (rs36228834), *CDKN2B* (rs2069416), *HDAC1* (rs1741981), *BRCA1* (rs3092986), *XRCC4* (rs3763063), and *XRCC5* (rs11685387) are associated with modified risk of disease, and pathway analysis supported a role for both biological processes in leukemogenesis (cell cycle OR(95%CI)= 1.21(1.07-1.38); DNA repair OR(95%CI)= 1.28(1.10-1.47)). This study warrants further elucidation of the mechanisms through which aberrant cell cycle control and double-strand break repair may be involved in childhood ALL susceptibility.

Introduction

B-cell precursor acute lymphoblastic leukemia (pre-B ALL) is the most common paediatric cancer. The origins of childhood ALL can likely be explained by a combination of inherited genetic predisposition and environmental exposure during early development in fetal life and infancy. The assertion that ALL may have a genetic basis has long been pursued through candidate gene association studies (1-5) and recently, genome-wide association studies (6-9) (GWAS) have vindicated the role of common inherited genetic variation in childhood ALL susceptibility; still few established genetic risk factors for ALL have been identified.

In this study, we postulated that a child's susceptibility to ALL may be influenced by a combination of sequence variants in multiple genes sharing functions in at least two core biological pathways: G1/S cell cycle checkpoint control and DNA double-strand break repair (DSBR). Strict regulation of the G1/S cell cycle checkpoint and concerted activation of DNA damage repair pathways is critical to maintain genomic integrity and cell homeostasis and protect against malignant transformation (10). Not surprisingly somatic inactivation of genes regulating DNA repair and cell cycle control are common features of most cancers in both adults and children (10-12). While there is well-established evidence for the role of these biological pathways in driving leukemogenesis (13, 14), their relevance in disease predisposition remains elusive. Because of the highly regulated nature of these pathways, it seems plausible that variation in gene expression levels due to regulatory polymorphisms, could perturb G1/S checkpoint activation and preclude proper DNA damage repair, leading to increased genomic instability and increased mutational burden resulting in leukemic transformation.

We hypothesized that *cis*-acting polymorphisms in the promoter regions (pSNPs) of genes encoding key players of the G1/S cell cycle checkpoint (*CCND1*, *CDC25A*, *CDKN1A*, *CDKN1B*, *CDKN2A*, *CDKN2B*, *E2F1*, *HDAC1*, *MDM2*, *SMAD3*, *RB1*, *TGFB1*) and DSB repair pathway (*ATM*, *BRCA1*, *BRCA2*, *RAD51*, *XRCC4*, *XRCC5*, *XRCC6*), could influence susceptibility to ALL. In an attempt to better understand the biological mechanisms underlying ALL predisposition, we applied a pathway-driven approach in which hierarchical modeling (HM) was used to perform a case-control association study and assess the effects of the above genes on childhood ALL risk either independently or collectively by assessing putative pathway-specific effects. An interesting feature of HM is that it allows prior information to be readily incorporated into the analytical framework, providing more accurate and stable risk estimates and reducing false-positive associations (15-17). Given that the promoter variants in our study were selected based on their putative impact on gene expression, we included prior information relevant to the functional regulatory potential of each variant such as *in silico* predictions regarding transcription factor (TF) binding as well as data from *in vitro* functional assays. This is the first study to directly incorporate prior biological and functional knowledge in a candidate gene association study on childhood ALL.

Material and Methods

Study subjects

Our cohort consisted of 321 childhood pre-B ALL patients and 329 healthy controls. Parental DNA was available for 203 of the probands. Study subjects were French-Canadians of European descent from the established Quebec Childhood ALL (QcALL) cohort (18). Incident cases were diagnosed in the Hematology-Oncology Unit of Sainte-Justine Hospital, Montreal, Canada,

between October 1985 and November 2006. The current study sample included 190 males and 131 females with a median age of 4.3 years. Healthy controls, 182 males and 146 females with a median age of 30.1 years, consisted of a group of newborns and adults recruited at Sainte-Justine Hospital. Peripheral blood or bone marrow (samples in remission) was collected from all participants and DNA was extracted as previously described (19). The Institutional Review Board approved the research protocol and informed consent was obtained from all participants and/or their parents.

SNP genotyping and quality control filters

A total of 46 promoter (defined as the 2kb region upstream of the transcription start site (20)) SNPs (pSNPs) in 19 genes were selected for genotyping (Table 1). SNPs were genotyped using the Luminex xMAP/Autoplex Analyser CS1000 system (Perkin Elmer, Waltham, MA). Groups of SNPs were amplified in multiplex PCR assays and hybridized to Luminex MicroPlex™-xTAG Microspheres for genotyping using allele-specific primer extension (ASPE) (21). The PCR and TAG-ASPE primers as well as amplification and reaction conditions are available upon request. Allele calls were assessed and compiled using the Automatic Luminex Genotyping software (M. Bourgey et al., manuscript under review). Hardy-Weinberg equilibrium (HWE) was tested in cases and controls separately using the chi-square goodness of fit test, and PedCheck (Version 1.1) was used to identify genotype incompatibilities using the familial data (22); inconsistent case-parent trios were sequenced for genotype validation and those that remained inconsistent were removed from the analysis.

Preliminary statistical analysis

Unless otherwise stated, statistical analyses were performed in STATA/IC version 10.1 (StataCorp, College Station, TX). Preliminary analysis was performed to dichotomize genotypes into two categories based on frequency distribution of the pSNP alleles, combining risk versus nonrisk genotypes. Logistic regression was used to compare genotype carriership in patients and controls and to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for all 46 pSNPs (Supplementary Table 1). The direction of the measured effect was used to identify risk genotype combinations defined as having $OR > 1$; therefore minor alleles were not necessarily assumed to confer an increased risk.

To reduce redundancy and overfitting in the regression models we chose a minimal subset of informative haplotype-tagging pSNPs for subsequent analyses. For genes with multiple markers, we reconstructed haplotypes using the expectation-maximization algorithm within FAMHAP (Version 16) (23), using parental data when available. The 'htsearch' program implemented in STATA (<http://www-gene.cimr.cam.ac.uk/clayton/software/>) with a pair-wise linkage disequilibrium threshold of $r^2 > 0.8$ was used to select informative tag pSNPs among cases and controls combined. Tag pSNPs and their associated high-risk genotypes are shown in Table 2.

Parental data and haplotype phase information, when available, were used to impute missing genotype data in the offspring. Only cases and controls with complete genotypic information at all loci (observed or imputed) were retained in subsequent analyses.

Hierarchical modeling

A two-stage hierarchical modeling (HM) approach was used to test for association between selected genes and childhood pre-B ALL. The hierarchical model was described previously (16, 17, 24). Briefly, we used a conventional logistic regression model as the first stage:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \mathbf{X}\boldsymbol{\beta} \quad (\text{A})$$

where \mathbf{X} is a matrix with row number equal to the number of individuals and column number equal to the number of genetic markers (n), and $\boldsymbol{\beta}$ is a column vector of regression coefficients representing the effects of each particular marker, on a log scale.

In the second-stage of the hierarchical model we used the following linear model to improve the estimation of the $\boldsymbol{\beta}$ coefficients:

$$\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\pi} + \boldsymbol{\delta} \quad (\text{B})$$

$$\delta_i \sim N(0, \tau^2)$$

where \mathbf{Z} is the prior covariate matrix with the i th row containing second-stage covariates (p) for the i th genetic marker, $\boldsymbol{\pi}$ is the corresponding column vector of p prior coefficients representing the linear effects of the second-stage covariates on β_i . The n -by- p \mathbf{Z} matrix reflects the similarities between the first-stage factors. The elements of $\boldsymbol{\delta}$ (δ_i) are the residual effects of the i genetic markers and are independent normal random variables with zero means and

variances τ^2 . The prior or second-stage variance, τ^2 , reflects the range of the potential residual effect that remains after accounting for all first- and second-stage covariates. In other words, the standard deviation, τ , is a measure of the prior uncertainty of β_i . For example $\tau = 0.5$ corresponds to being 95% certain a priori that the OR of a given genetic marker lies within a 7-fold range ($\exp(3.92\tau) = \exp(3.92*0.5) = 7$) (15).

Substituting equation B into A gives the following mixed model:

$$\begin{aligned} \ln\left(\frac{P}{1-P}\right) &= \alpha + \mathbf{X}(\mathbf{Z}\underline{\pi} + \underline{\delta}) \\ &= \alpha + \mathbf{XZ}\underline{\pi} + \mathbf{X}\underline{\delta} \end{aligned} \quad (C)$$

in which $\underline{\pi}$ is a vector of fixed coefficients and $\underline{\delta}$ is treated as a vector of random coefficients with mean 0 and variance $\underline{\tau}^2$. Conventional maximum-likelihood estimation is equivalent to using a model (C) with no constraint on the residual effects ($\underline{\tau}^2 = \infty$) and could result in overfitting of the data, conversely, $\underline{\tau}^2 = 0$ implies no uncertainty about the residual effects and can lead to underdispersion of the HM estimates, compared to the constrained ($\underline{\tau}^2 < \infty$) multilevel approach (25).

Specification of the second-stage covariates

We constructed the \mathbf{Z} matrix using prior information pertaining to the biological function of the gene and the potential significance of its pSNPs (Table 2). First, an indicator column was created to represent each biological pathway (G1/S cell cycle checkpoint and DSBR) and a score of 1 was assigned to the genetic risk

variant if the corresponding gene was involved in the pathway or 0 if it was not; we assumed that the pathways were mutually exclusive. Genes assigned to a similar pathway were therefore assumed to be interchangeable and their effects to arise from a common distribution.

Second, we included prior information relevant to the functional regulatory impact of each polymorphism on gene expression. The putative effect of each pSNP on transcription factor (TF) binding was evaluated through *cis*-regulatory module (CRM) predictions (26-28). *In silico* analysis was performed using the PreMod database (<http://genomequebec.mcgill.ca/PReMod>) to evaluate the presence of CRMs; genomic regions with significant module scores (P -values $< 5 \times 10^{-5}$) reported in PReMod reflect the presence, in a region of 100-1000bp, of a large number of conserved TF binding sites (29). pSNPs falling in such regions were assigned the corresponding module score in the **Z** matrix; a score of 0 was assigned if no significant CRM was identified in the 1kb region surrounding the polymorphism. *In vitro* analyses were carried out to assess the functional significance of the variants on DNA-protein binding using electrophoretic mobility shift assays (EMSA), and on allele-specific promoter activity using luciferase gene reporter assays. Experimental validation was performed in three different cell lines (HepG2, Jeg-3 and HeLa) as described previously (30). A score of 1 was assigned to a variant in the **Z** matrix if the EMSA showed appearance and/or disappearance of a DNA-protein complex in the presence of the risk allele in at least one of the three cell lines while a score of 0 was assigned if there was no change in DNA-protein complex formation. Similarly, a score of 1 was assigned to a variant if a significant effect on promoter activity (increase or decrease of reporter gene expression) was observed for the risk allele/haplotype compared to the non-risk allele/haplotype in at least one of the cell lines tested; a score of 0 was assigned if no change in reporter gene expression was observed. To include the maximal amount of prior information in the analysis we used a conservative approach and assumed that

variants with missing data had no relevant functional significance, in which case a score of 0 was assigned to the corresponding Z matrix column for that variant.

The resulting second-stage Z matrix had 5 columns for each of the prior covariates (DSBR pathway, cell cycle pathway, CRM score, differential DNA-protein binding, and allele-specific promoter activity) (Table 2). Two variations of this prior matrix were used in the HM analysis: a reduced model including only the two pathway indicators and the full model including both the pathway indicators and functional covariates.

Specification of the second-stage variance

We performed sensitivity analysis to select an appropriate range of residual main effects, τ^2 . Hierarchical modeling was performed with pathway indicators only and τ^2 set to 0.01, 0.04, 0.1225, 0.25, 0.4 or 1 which corresponds to being 95% certain a priori that an OR lies in a 1.5-fold, 2.2-fold, 4-fold, 7-fold, 12-fold and 50-fold range, respectively. We set all τ^2 to a single value, implying equal uncertainty about all residual effects. Given that the second-stage model contained only limited prior information on the biological relevance of the genes and the functional significance of their polymorphisms, the existence of a large residual effect is theoretically possible. However most common susceptibility variants identified so far have only low to intermediate effects on risk (31), therefore it seems unlikely that any of the variants under investigation would confer a particularly large increment in disease risk. Consequently, for association testing of main genetic effects we set τ^2 to a moderate value of 0.04 which would allow a 2.2-fold variation in the effect of the gene variant. Nonetheless, even when τ^2 was as large as 1 (implying a prior 95% CI covering a 50-fold range) results were similar (Supplementary Table 2).

Association testing using hierarchical modeling and conventional logistic regression

We used the SAS macro GLIMMIX for multilevel modeling (SAS version 9.00, SAS Institute Inc., Cary, NC) in conjunction with the SAS IML matrix language (source code provided by Witte et al. (25); http://www.epibiostat.ucsf.edu/witte_lab/glimmix.htm) for two-stage hierarchical modeling of the childhood pre-B ALL case-control data.

To evaluate the performance of our model, we compared the results based on HM to those from conventional maximum-likelihood analysis. We either treated all genetic factors independently and constructed logistic regression models with a single marker in each regression (Conventional model 1); or included all markers in one single logistic regression model (Conventional model 2). Pathway effects were further validated using the pathway genetic load (PGL) method described in Huebinger et al. (32). Individuals were assigned a score based on the number of risk genotypes they carried across all loci of a given pathway. Scores for the genetic load for the cell cycle pathway could range from 0 to 20 and from 0 to 14 for the DSB pathway. Logistic regression was used, with PGL as independent variable, to measure the risk associated with carrying an increasing number of risk genotypes in a given pathway.

Gene-gene interaction effects

To explore possible epistasis effects in the hierarchical modeling framework, we added product terms of pair-wise interactions to the first-stage model of the form:

$$\ln\left(\frac{P}{1-P}\right) = \alpha + X_i\beta_i + X_j\beta_j + (X_i * X_j)\beta_{ij} \quad (D)$$

An extra row was added to the Z matrix for each gene-gene product term and each interaction parameter was assigned a score of 1 in the relevant biological pathway columns. Given that the functional data was relevant to each individual pSNP and no measures of the putative biological effects of statistical interactions between pSNPs on disease were available, we included only pathway information in the gene-gene interaction analysis; the Z matrix is available upon request.

Consistent with the main effects, we performed a sensitivity analysis to select an appropriate range of residual gene-gene interaction effects (Supplementary Table 3). We selected a slightly wider range and set $\tau^2 = 0.1225$ allowing a 4-fold variation in the effect of the gene x gene product term on childhood B-cell ALL risk.

The SAS macro GLIMMIX was used to test for gene-gene interactions in the HM framework, as described previously. We compared the HM estimates to those obtained through conventional gene-gene interaction analyses. We performed maximum-likelihood analysis using logistic regression with either a single gene x gene (or effectively a SNP x SNP) product term and both significant main effects in each regression model (Conventional GxG model 1), or all significant main effects and all relevant pair-wise SNP x SNP product terms in one single model (Conventional GxG model 2). We performed multiple logistic regressions in a backward stepwise manner (Backward stepwise model) forcing all significant main effects into the model and using a significance level for removal from the model of $P \leq 0.05$ and a significance level for addition to the model of $P \leq 0.10$.

Results

A total of 46 pSNPs in 19 genes were genotyped among 321 unrelated cases (including 203 case-parent trios) and 329 controls. All of the SNP allele frequencies were in HWE ($P < 0.01$) except for variants *CCND1*-1537INS/C and *XRCC6*-1469C>T which significantly deviated from HWE in the control group and were therefore excluded from the analysis. We also removed individuals for which Mendelian inconsistencies could not be reconciled by sequencing; on average, two families were removed at each locus. The percentage of missing case genotypes was well below 10% for all variants except *ATM*-635T>A (12.1%). Following imputation using parental data and LD information, the percentage of missing data dropped to less than 3.4% across all loci. Imputation had no effect on the significance or directionality of the associations therefore *ATM*-635T>A was retained in the analysis. Variant genotypes were dichotomized as risk versus nonrisk, assuming that the risk genotype combination was associated with a positive effect ($OR > 1$) on disease susceptibility (Supplementary Table 1). Following data quality control and imputation a total of 286 childhood pre-B ALL cases and 297 healthy controls were fully genotyped at 34 tag-pSNPs from 19 genes (Table 2). The distribution of demographic and clinical characteristics of cases retained in the analysis are given in Table 3.

We first measured the main effects of the pSNPs on pre-B ALL risk (Table 4). For each variant, four risk estimates ($OR(95\%CI)$) are shown. The conventional single marker analysis suggested that seven pSNPs were associated with pre-B ALL risk when considered independently: *CDKN2A*-222T>A, *CDKN2B*-593A>T,C, *TGFB1*-1886A>G, *BRCA1*-588A>G, *XRCC4*-1864T>C, *XRCC4*-1407C>G, and *XRCC5*-297C>T. Association with three of these variants (*TGFB1*-1886A>G, *XRCC4*-1407C>G, and *XRCC5*-297C>T) was lost using

multivariate analysis. The HM procedure with pathway indicators only, confirmed the association of variants *CDKN2A*-222T>A, *CDKN2B*-593A>T,C, *BRCA1*-588A>G, *XRCC4*-1864T>C, and *XRCC5*-297C>T and an additional effect was observed for *HDAC1*-1269T>C, but the latter was lost when we included functional data in the prior covariate matrix. Hierarchical regression increased the precision of the risk estimates as CIs were narrower when compared to conventional estimates, and estimates were more stable: more extreme values in the conventional analysis, such as the risk estimate for *CDKN2A*-222TA/AA (Conventional model 1: OR(95%CI)= 2.38(1.34-4.22)), experienced notable shrinkage (Hierarchical model 1: OR(95%CI)= 1.40(1.00-1.96)); moderate, stable risk estimates however, such as *XRCC4*-1864T>C (Conventional model 1: OR(95%CI)= 1.61(1.13-2.30)) remained much the same ((Hierarchical model 1: OR(95%CI)= 1.42(1.07-1.88)). Results for Hierarchical model 2 (pathway indicators and functional assays) were more unstable and had larger variances than those obtained including pathway information only. This may be due to the fact that functional information was lacking for several of the pSNPs.

Rather than limit the analysis to individual genes, we investigated the combined action of multiple genes within the same pathway. Assuming that the effects of the SNPs from a similar pathway are interchangeable and are drawn from the same distribution, we could extract the second-stage coefficients (π) from the HM model to measure pathway-specific effects on disease outcome. Both the DSB and G1/S cell cycle control pathways were associated with increased risk of pre-B ALL among children (Hierarchical model 1: OR(95%CI)= 1.28(1.10-1.47) and 1.21(1.07-1.38), respectively) (Table 4). Similar results were obtained when the pSNPs were combined into a pathway genetic load score; significant associations were confirmed with both the cell cycle control (OR(95%CI)= 1.19(1.10-1.29), $P=2.10 \times 10^{-5}$) and DSB pathways (OR(95%CI)= 1.27(1.15-

1.39), $P=3.61 \times 10^{-7}$), and increased risk of ALL was associated with increasing numbers of risk genotypes within a given pathway.

Finally, we investigated the putative effects of gene-gene interactions on pre-B ALL risk using both conventional maximum-likelihood estimation and hierarchical modeling (Supplementary Table 4). A fully saturated model including all main effects and all possible pair-wise interactions resulted in model instability due to sparse data. Therefore we limited the gene-gene interaction analysis to variants with significant main effects only. Three conventional maximum-likelihood testing procedures for detecting gene-gene interactions were compared to HM in which product terms for the 15 pair-wise interactions were included in the first-stage regression model and pathway indicators informed the second-stage regression. HM provided overall more stable values, however no evidence of interaction could be detected in this dataset. We also tested all possible pair-wise interaction effects between the 34 pSNPs using conventional logistic regression (561 tests in total); as expected, none of the significant interactions withstood multiple testing corrections (data not shown). In addition, we modeled pathway-specific interactions by restricting the analysis to genes acting within the same pathway, since these are likely to be biologically more relevant, yet no further evidence of interaction was observed (data not shown).

Discussion

Using a comprehensive hierarchical modeling approach, we incorporated biological knowledge as well as *in silico* and *in vitro* functional data into the association analysis, and successfully identified individual variants in two biologically relevant pathways that were associated with an increased risk of

ALL: *CDKN2A*-222TA/AA (rs36228834), *CDKN2B*-593AA/AC (rs2069416), and *HDAC1*-1269TT (rs1741981) involved in the G1/S cell cycle checkpoint, and *BRCA1*-588AA (rs3092986), *XRCC4*-1864TT (rs3763063), and *XRCC5*-297CT/TT (rs11685387) involved in DNA double-strand break repair.

CDKN2A and *CDKN2B* are well-characterized tumor suppressors that map to chromosomal region 9p21. They negatively regulate cell cycle progression through the G1/S checkpoint and are frequently inactivated in B-cell ALL (33). The implication of these genetic variants in ALL susceptibility has previously been demonstrated in a candidate-gene study on the same cohort (5). Furthermore, a recent GWAS identified a variant (rs3731217) within the *CDKN2A* gene region that was strongly associated with childhood ALL in a large case-control replication series including French-Canadian patients from the QcALL cohort (8). This variant lies 9.3kb downstream from the *CDKN2A* association signal identified here. Unfortunately we are currently unable to describe LD levels between these two SNPs therefore it is unknown to us whether or not these *CDKN2A* signals represent the same association. However LD between variants *CDKN2A*-222T>A and *CDKN2B*-593A>T,C in the present study is low ($D' = 0.81$, $R^2 = 0.02$), and based on data from European individuals genotyped in the 1000 Genomes Project, pairwise LD between variants *CDKN2A* rs3731217 from the GWAS and *CDKN2B*-593A>T,C from this study, is also expected to be low ($D' = 0.32$ and $R^2 = 0.01$), suggesting independence of the *CDKN2A* and *CDKN2B* association signals. While Sherborne et al. found no SNPs in LD with variant rs3731217 in any of the coding regions of *CDKN2A* or in any of the other adjacent genes (*CDKN2B*, *MTAP*, *ANRIL*), nor did they identify any regulatory regions surrounding rs3731217 or effects on gene expression (8), our *in vitro* data did show differential DNA-protein binding within the promoter region of *CDKN2A* and allele-specific promoter activity associated with the variant -222T>A (rs36228834) in our study. Additional functional data including allele-specific promoter activity associated with promoter haplotypes

carrying the risk alleles for *CDKN2A*-222A and *CDKN2B*-593A, in conjunction with the differential DNA-protein binding observed for allele *CDKN2A*-222A (no complex-binding differences were observed for *CDKN2B*-593A>T,C), further support the putative role of these variants in modulating gene expression and perhaps disease risk. Yet fine-mapping and additional functional studies are imperative to better define the causal basis of the 9p21 associations with childhood pre-B ALL.

Furthermore, histone deacetylase *HDAC1* plays a key role in lineage specification during hematopoiesis and is overexpressed in acute myeloid leukemia where it has been shown to contribute to leukemogenesis by perturbing differentiation (34). In our gene reporter assays, variant -1269T in *HDAC1* was associated with increased promoter activity in HeLa cells (no EMSAs were available). *BRCA1* is a master regulator of genome integrity due to its roles in both cell cycle checkpoint control and DSBR through homologous recombination (35). Inactivation of *BRCA1* has been shown to increase risk for leukemia by allowing the number of cells with errors in DSBR to survive (36); in fact *BRCA1* is nearly undetectable in leukemia cells from patients with chronic myelocytic leukemia (37). The *in silico* analysis was highly indicative of a CRM within the 1kb surrounding the *BRCA1*-588A>G variant and functional validation showed differential DNA-protein complex formation and gene reporter assays suggested allele-specific gene expression associated with promoter haplotypes carrying the *BRCA1*-588A allele. Conversely, *XRCC4* and *XRCC5* are involved in the error-prone nonhomologous end joining repair of double strand breaks. Overactivity of this pathway and concomitant DNA misrepair has been shown to lead to increased chromosomal instability and contribute to leukemia (38). Though no EMSAs were available, luciferase gene reporter assays did provide evidence for allele-specific promoter activity associated with pSNP *XRCC5* -297C>T with the high-risk -297T allele showing increased activity *in vitro*. Promoter SNPs in *XRCC4* have recently been associated with increased

childhood leukemia risk in a Taiwanese population (39); yet unfortunately no functional data were available for variant *XRCC4* -1864.

Our pathway-based analysis further supported a role for the G1/S cell cycle checkpoint and DSBR systems in ALL susceptibility perhaps via increased tolerance to genomic instability associated with impaired differentiation and deregulated proliferation. Risk was shown to significantly increase with increasing numbers of risk alleles within a given pathway, as demonstrated by the PGL analysis. Thus, based on the observed associations and *in vitro* functional evidence, one could speculate that modulated expression levels of G1/S cell cycle regulators *CDKN2A*, *CDKN2B* and *HDAC1*, and of DSBR genes *BRCA1*, *XRCC4* and *XRCC5*, due to the corresponding promoter polymorphisms, could influence a child's susceptibility to develop ALL either through the individual effects of their variants or via global pathway deregulation. However the functional significance of the observed associations remains speculative and requires further validation to confirm association and better describe the mechanistic role of these pSNPs and respective pathways in childhood pre-B ALL susceptibility.

This is the first time HM was applied in the study of childhood leukemia, and to the best of our knowledge, this is the first time functional data have efficiently been integrated in a case-control genetic association study to inform risk estimation. HM allowed us to increase the precision of the risk estimates and reduced the likelihood of false-positives by shrinking effect sizes toward a prior mean. Though a more comprehensive modeling approach within both the conventional and hierarchical framework would include main effects and putative gene-gene interactions between factors that act jointly within a pathway or that operate in interconnecting pathways, we were unable to detect any significant gene-gene interaction effects in this study, which suffered from small

sample sizes. Still epistasis likely plays an important role in determining disease predisposition and warrants further investigation in larger childhood pre-B ALL study samples. Additional limitations to the hierarchical model used in this study include the dichotomization of the SNP genotypes into risk versus nonrisk classes, which may not have reflected the underlying genetic model of inheritance. Second, in the design of the \mathbf{Z} matrix, prior functional information was missing for some of the variants and was considered equivalent to a negative result. Moreover, use of 0 and 1 in the \mathbf{Z} matrix to describe EMSA and gene reporter data, is unlikely to capture the precise differences between the genetic factors. Third, in using a single common prior variance, τ^2 , for all markers, we assumed that the effects of individual SNPs and gene-gene interactions from a similar pathway were drawn from a single distribution and that their effect sizes should be shrunk toward a single prior mean; however in practice, this may not hold true. Nonetheless, our results suggest that the pathway-based model was relatively robust and that a \mathbf{Z} matrix that is reasonable in its biological interpretation of the functional significance of genetic variants can be helpful in improving risk estimation. Finally, even though the HM model used in this study obviously represented an incomplete picture of the underlying disease process, it still provided insight into the mechanisms that may be driving leukemogenesis, suggesting a role for both the cell cycle and DNA double-strand break repair pathways, and allowed for individual gene variants associated with disease to be identified.

In conclusion, using a comprehensive hierarchical modeling approach, we provide evidence that genes involved in mediating the cellular response to DNA damage, including G1/S cell cycle inhibition and DSBR, are associated with modified risk of pre-B ALL among children. However, replication in larger cohorts is imperative and further association testing and functional validation will also be required to identify the causal variants at the suggested loci and validate their implication in disease. This is especially true given that tag SNPs were

used in the analysis and that these SNPs are in linkage disequilibrium with a number of SNPs that may or may not have been genotyped in this study. Considering that new therapeutic development in cancer may lie in the discovery of agents that target the physiologic effects of altered biologic pathways rather than their individual gene components (40), better elucidation of the mechanisms through which aberrant cell cycle control and DSBs are involved in childhood ALL susceptibility could be of great diagnostic value and provide the data required to help guide risk-directed therapy, ultimately improving disease management and outcome.

Tables

Table 1. Genes and DNA variants genotyped in B-cell ALL patients, their parents, and healthy controls

Pathway: Gene (Chr.), DNA variant	dbSNP	Position	MAF (%)
G1/S cell cycle control:			
<i>CCND1</i> (11q13)			
-1938T>C	rs1944129	69,453,935	48.8
-1537INS C	rs36225395	69,454,336	45.2
<i>CDC25A</i> (3p21)			
-2030G>T	rs1903061	48,231,919	10.3
<i>CDKN1A</i> (6p21.2)			
-1284T>C	rs733590	36,645,203	36.2
-899T>G	rs762624	36,645,588	27.1
-791T>C	rs2395655	36,645,696	39.74
<i>CDKN1B</i> (12p13)			
-1857C>T	rs3759217	12,868,452	11.6
-1608G>A	rs35756741	12,868,701	8.6
-373G>T	rs36228499	12,869,936	43.4
<i>CDKN2A</i> (9p21)			
-222T>A	rs36228834	21,975,319	3.3
<i>CDKN2B</i> (9p21)			
-1270C>T	rs36229158	22,010,681	2.8
-593A>T,C	rs2069416	22,010,004	37.4/2.7
-287G>C	rs2069418	22,009,698	42.7
<i>E2F1</i> (20q11.2)			
-187C>T	rs3213141	32,274,380	24.1
<i>HDAC1</i> (1p34.1)			
-1269T>C	rs1741981	32,756,439	33.0
-455T>C	rs36212119	32,757,253	8.4
<i>MDM2</i> (12q14.3-q15)			
-1494A>G	rs1144944	69,200,485	49.2
-1174DEL AAAAAGC(40bp)	rs3730485	69,200,806-69,200,845	40.5
-182C>G	rs937282	69,201,797	48.3
+309T>G	rs2279744	69,202,580	36.6
<i>SMAD3</i> (15q21-q22)			
-1938T>C	rs36221701	67,356,489	12.8
<i>RB1</i> (13q14.1-q14.2)			
-1554C>A	rs1573601	48,876,357	24.8
<i>TGFB1</i> (19q13.1)			
-1886A>G	rs2317130	41,861,674	31.4
-1571G>A	rs4803457	41,861,359	39.4
-1550DEL AGG	rs11466313	41,861,338-41,861,337	31.0
-508G>A	rs1800469	41,860,296	31.3

DNA double-strand break repair:				
<i>ATM</i> (11q22.3)				
-1206G>T	rs4987876	108,092,637		9.6
-635T>A	rs228589	108,093,208		41.8
<i>BRCA1</i> (17q21)				
-1890T>C	rs4793204	41,279,298		32.4
-708A>G	rs799906	41,278,116		33.4
-598INS ACA	rs8176071	41,278,006-41,278,005		32.3
-588A>G	rs3092986	41,277,996		9.0
<i>BRCA2</i> (13q12.3)				
-1555C>A	rs206114	32,888,062		41.3
-1260DEL GTCTAA	rs3072036	32,888,357-32,888,362		39.6
-1144A>G	rs206115	32,888,473		40.9
-1134C>T	rs206116	32,888,483		40.2
-908C>T	rs206117	32,888,709		39.7
-254G>A	rs3092989	32,889,363		20.5
<i>RAD51</i> (15q15.1)				
-1185A>T	rs2619679	40,986,237		49.1
<i>XRCC4</i> (5q13-q14)				
-1864T>C	rs3763063	82,371,453		49.5
-1407C>G	rs1993947	82,371,910		8.4
<i>XRCC5</i> (2q35)				
-1379G>T	rs828907	216,972,732		47.1
-297C>T	rs11685387	216,973,814		25.6
<i>XRCC6</i> (22q11-q13)				
-1469C>T	rs28384701	42,016,526		4.5
-1296C>G	rs2267437	42,016,699		38.6
-731G>A	rs132770	42,017,264		25.1

DNA variant positions were numbered with respect to the first nucleotide of the first exon as +1, and the nucleotide immediately upstream as -1. Mapping of SNP positions is based on dbSNP build 130 and the GRCh37 human assembly (UCSC Genome Browser). Minor allele frequency (MAF) indicates minor allele frequency and was calculated on a control cohort consisting of 329 healthy French-Canadian individuals from the QcALL cohort; overall frequencies are comparable to those reported for other populations of European descent. SNPs that were included in the analysis had a $MAF \geq 5\%$ in at least the pre-B ALL patients or the control group.

Table 2. Prior matrix used in the hierarchical model

Gene	tag pSNP	High-risk genotype	DSBR	Cell cycle	CRM score	Differential DNA-protein binding	Allele-specific promoter activity
<i>CCND1</i>	-1938T>C	-1938CC	0	1	33.83	0	0
<i>CDC25A</i>	-2030G>T	-2030GG	0	1	0	1	1
<i>CDKN1A</i>	-1284T>C	-1284TC/CC	0	1	0	1	1
	-899T>G	-899TT	0	1	0	1	1
<i>CDKN1B</i>	-1857C>T	-1857CC	0	1	0	0	0
	-1608G>A	-1608GA/AA	0	1	0	0	0
	-373G>T	-373GG/GT	0	1	38.14	0	1
<i>CDKN2A</i>	-222T>A	-222TA/AA	0	1	0	1	1
<i>CDKN2B</i>	-593A>T,C	-593AA/AC	0	1	0	0	1
	-287G>C	-287GG	0	1	0	1	1
<i>E2F1</i>	-187C>T	-187CC	0	1	0	0	0
<i>HDAC1</i>	-1269T>C	-1269TT	0	1	0	0	1
	-455T>C	-455TC/CC	0	1	0	0	1
<i>SMAD3</i>	-1938T>C	-1938TT	0	1	0	0	1
<i>MDM2</i>	-1494A>G	-1494AG/GG	0	1	0	1	1
	-1174DEL AAAAAGC(40bp)	-1174AAAAAGC/--	0	1	0	0	1
	+309T>G	+309TT	0	1	0	0	1
<i>RB1</i>	-1554C>A	-1554CC	0	1	0	1	0
<i>TGFB1</i>	-1886A>G	-1886GG	0	1	0	1	1
	-1571G>A	-1571AA	0	1	0	0	1
<i>ATM</i>	-1206G>T	-1206GT/TT	1	0	25.37	0	1
	-635T>A	-635TT	1	0	25.37	0	1

<i>BRCA1</i>	-598INS ACA	-598-ACA/ACA ACA	1	0	56.67	1	1
	-588A>G	-588AA	1	0	56.67	1	1
<i>BRCA2</i>	-1260DEL GTCTAA	-1260--	1	0	0	0	1
	-1144A>G	-1144GG	1	0	0	1	1
	-254G>A	-254GG	1	0	0	0	1
<i>RAD51</i>	-1185A>T	-1185TT	1	0	0	0	0
<i>XRCC4</i>	-1864T>C	-1864TT	1	0	0	0	0
	-1407C>G	-1407CG/GG	1	0	0	0	0
<i>XRCC5</i>	-1379G>T	-1379GG	1	0	0	0	1
	-297C>T	-297CT/TT	1	0	0	0	1
<i>XRCC6</i>	-1296C>G	-1296GG	1	0	0	0	0
	-731G>A	-731GG	1	0	0	0	0

The **Z** matrix used in the second stage of the hierarchical model included pathway information (DSBR double strand break repair; cell cycle control) as well as prior information on potential pSNP function based on *cis*-regulatory modulation potential (CRM score), on disruption of DNA-protein binding as evidenced by electrophoretic mobility shift assays (EMSA), and on allele-specific promoter activity as evidenced by gene reporter assays. Results for the EMSA and reporter gene expression assay were dichotomized: 1= appearance and/or disappearance of DNA-protein complex in the presence of the risk allele or significant increase or decrease of gene expression for the risk allele/haplotype compared to the non-risk allele/haplotype; 0= no change in DNA-protein complex formation in the presence of the risk allele or no change in reporter gene expression for the risk allele/haplotype. High risk alleles were presumed to have a positive effect on disease risk and therefore minor alleles were not necessarily assumed to confer an increased risk.

Table 3. Characteristics of the French-Canadian pre-B ALL patients from the Quebec Childhood ALL cohort that were retained in the analysis

Patient Characteristics	Cases, n (%)
Total number of subjects	286
Gender	
Male	168 (58.7)
Female	118 (41.3)
Age group, years	
≤ 1	6 (2.1)
1-10	220 (76.92)
> 10	42 (14.69)
N/D	18 (6.29)
Hyperdiploidy	
Positive	109 (38.1)
Negative	150 (52.5)
N/D	27 (9.4)
Chromosomal translocations	
Absence of translocation	92 (32.2)
t(12;21)	32 (11.2)
Other	10 (3.4)
N/D	152 (53.2)
Normal ploidy and absence of chromosomal translocations	39 (13.6)

Frequencies based on the total number patients that were fully genotyped at all tag-pSNPs following data quality control and imputation. N/D indicates no data available.

Table 4. Main effects of pSNPs on pre-B ALL risk among children, as estimated by conventional maximum-likelihood analysis and hierarchical modeling

Gene	tag pSNP	High-risk genotype	Case no. (n= 286)	Control no. (n= 297)	Conventional 1, OR (95% CI)	Conventional 2, OR (95% CI)	Pathway HM 1, OR (95% CI)	Pathway and functional HM 2, OR (95% CI)
<i>CCND1</i>	-1938T>C	-1938CC	67	60	1.21 (.82-1.79)	1.26 (0.82-1.93)	1.23 (.92-1.66)	1.24 (.88-1.74)
<i>CDC25A</i>	-2030G>T	-2030GG	237	239	1.17 (.77-1.79)	1.24 (.78-1.96)	1.24 (.91-1.69)	1.27 (.92-1.75)
<i>CDKN1A</i>	-1284T>C	-1284TC/CC	181	178	1.15 (.83-1.61)	1.28 (.85-1.93)	1.23 (.93-1.63)	1.26 (.94-1.69)
	-899T>G	-899TT	160	161	1.07 (.77-1.49)	1.22 (.81-1.82)	1.21 (.91-1.60)	1.24 (.92-1.66)
<i>CDKN1B</i>	-1857C>T	-1857CC	228	229	1.17 (.79-1.73)	1.16 (.74-1.83)	1.23 (.91-1.65)	1.21 (.89-1.65)
	-1608G>A	-1608GA/AA	56	50	1.20 (.79-1.83)	1.14 (.71-1.84)	1.17 (.87-1.59)	1.16 (.85-1.59)
	-373G>T	-373GG/GT	241	247	1.08 (.70-1.68)	1.04 (.63-1.71)	1.14 (.84-1.56)	1.15 (.81-1.63)
<i>CDKN2A</i>	-222T>A	-222TA/AA	40	19	2.38 (1.34-4.22)	1.98 (1.05-3.72)	1.40 (1.00-1.96)	1.44 (1.01-2.04)
<i>CDKN2B</i>	-593A>T,C	-593AA/AC	136	113	1.48 (1.06-2.05)	1.48 (1.01-2.18)	1.38 (1.05-1.81)	1.36 (1.03-1.79)
	-287G>C	-287GG	96	94	1.09 (.77-1.54)	1.20 (.80-1.81)	1.23 (.93-1.63)	1.25 (.93-1.67)
<i>E2F1</i>	-187C>T	-187CC	166	170	1.03 (.74-1.44)	1.03 (.72-1.47)	1.13 (.86-1.47)	1.12 (.85-1.47)
<i>HDAC1</i>	-1269T>C	-1269TT	147	131	1.34 (.97-1.86)	1.43 (.99-2.06)	1.31 (1.00-1.71)	1.30 (.99-1.70)
	-455T>C	-455TC/CC	54	48	1.21 (.79-1.85)	1.18 (.74-1.89)	1.19 (.88-1.62)	1.18 (.86-1.61)
<i>SMAD3</i>	-1938T>C	-1938TT	224	226	1.14 (.77-1.67)	1.17 (.77-1.79)	1.18 (.88-1.58)	1.17 (.87-1.57)
<i>MDM2</i>	-1494A>G	-1494AG/GG	224	226	1.14 (.77-1.67)	0.95 (.51-1.80)	1.08 (.79-1.48)	1.12 (.80-1.56)
	-1174DEL AAAAAGC(40bp)	-1174AAAAAGC/--	191	190	1.13 (.80-1.59)	1.06 (.60-1.87)	1.09 (.81-1.46)	1.07 (.79-1.44)
	+309T>G	+309TT	127	111	1.34 (.96-1.86)	1.34 (.89-1.94)	1.26 (.96-1.65)	1.24 (.94-1.64)
<i>RB1</i>	-1554C>A	-1554CC	162	161	1.10 (.80-1.53)	1.15 (.81-1.64)	1.18 (.90-1.54)	1.20 (.90-1.61)
<i>TGFB1</i>	-1886A>G	-1886GG	37	23	1.77 (1.02-3.06)	2.57 (1.00-6.56)	1.32 (.93-1.85)	1.36 (.95-1.96)
	-1571G>A	-1571AA	48	41	1.26 (.80-1.98)	0.69 (.31-1.53)	1.15 (.83-1.59)	1.12 (.80-1.57)
<i>ATM</i>	-1206G>T	-1206GT/TT	52	52	1.05 (.69-1.60)	0.96 (.60-1.54)	1.15 (.84-1.56)	1.14 (.83-1.58)
	-635T>A	-635TT	121	109	1.26 (.91-1.76)	1.33 (.91-1.93)	1.27 (.97-1.67)	1.28 (.97-1.69)
<i>BRCA1</i>	-598INS ACA	-598-ACA/ACA ACA	164	158	1.18 (.85-1.64)	1.12 (.78-1.61)	1.21 (.93-1.59)	1.23 (.93-1.63)
	-588A>G	-588AA	256	246	1.77 (1.09-2.87)	1.84 (1.10-3.10)	1.45 (1.05-2.00)	1.54 (1.02-2.32)
<i>BRCA2</i>	-1260DEL GTCTAA	-1260--	59	47	1.38 (.91-2.11)	1.47 (.88-2.46)	1.33 (.97-1.82)	1.32 (.96-1.83)
	-1144A>G	-1144GG	63	51	1.36 (.90-2.05)	1.38 (.84-2.28)	1.29 (.95-1.76)	1.33 (.96-1.85)
	-254G>A	-254GG	199	187	1.35 (.95-1.90)	1.10 (.72-1.70)	1.21 (.92-1.60)	1.20 (.90-1.59)

<i>RAD51</i>	-1185A>T	-1185TT	71	68	1.11 (.76-1.63)	1.11 (.73-1.67)	1.17 (.88-1.57)	1.17 (.87-1.57)
<i>XRCC4</i>	-1864T>C	-1864TT	102	76	1.61 (1.13-2.30)	1.53 (1.04-2.27)	1.42 (1.07-1.88)	1.41 (1.06-1.87)
	-1407C>G	-1407CG/GG	67	50	1.51 (1.00-2.27)	1.31 (.83-2.05)	1.30 (.96-1.76)	1.29 (.95-1.75)
<i>XRCC5</i>	-1379G>T	-1379GG	91	78	1.31 (.91-1.88)	1.17 (.78-1.78)	1.23 (.92-1.63)	1.22 (.92-1.63)
	-297C>T	-297CT/TT	158	136	1.46 (1.05-2.02)	1.39 (.96-2.01)	1.33 (1.02-1.74)	1.32 (1.00-1.73)
<i>XRCC6</i>	-1296C>G	-1296GG	50	39	1.40 (.89-2.21)	1.37 (.81-2.32)	1.30 (.95-1.78)	1.29 (.93-1.78)
	-731G>A	-731GG	178	170	1.23 (.88-1.71)	1.17 (.79-1.71)	1.22 (.93-1.60)	1.21 (.92-1.59)
Pathway estimation								
DSBR								1.28 (1.10-1.47) 1.25 (1.05-1.50)
Cell cycle								1.21 (1.07-1.38) 1.19 (.97-1.44)

Conventional model 1, logistic regression with a single marker in each regression; Conventional model 2, logistic regression with all markers in one single model; Pathway HM 1, hierarchical modeling with τ^2 set to 0.04 and pathway indicators only; Pathway and functional HM 2, hierarchical modeling with τ^2 set to 0.04, pathway indicators and functional covariates including CRM scores, as well as indicator variables for differential DNA-protein binding and allele-specific gene expression. Significant results are shown in bold.

HM, hierarchical modeling; OR, odds ratio; CI, confidence interval; DSBR, double strand break repair.

Supplementary Table 1. Distribution of DNA repair and cell cycle genotypes among B-cell ALL cases and controls from the Quebec Childhood ALL cohort and their effect on ALL risk, as estimated by logistic regression

Pathway: Gene, DNA variant	db SNP	Genotype	No. (%)		OR (95% CI)	P
			ALL patients	Controls		
G1/S cell cycle control:						
CCND1						
-1938T>C	rs1944129	TT	78 (25.1)	74 (23.0)	1 (referent)	—
		TC	162 (52.1)	182 (56.5)	0.84 (.58-1.24)	.39
		CC	71 (22.8)	66 (20.5)	1.02 (.64-1.62)	.93
-1537INS C	rs36225395	-/-	85 (27.3)	85 (26.2)	1 (referent)	—
		-/C	163 (52.4)	186 (57.2)	0.88 (.61-1.26)	.48
		CC	63 (20.3)	54 (16.6)	1.17 (.73-1.87)	.52
CDC25A						
-2030G>T	rs1903061	GG	253 (83.5)	262 (80.4)	1 (referent)	—
		GT	47 (15.5)	61 (18.7)	0.80 (.52-1.21)	.29
		TT	3 (1.0)	3 (0.9)	1.04 (.21-5.18)	.97
CDKN1A						
-1284T>C	rs733590	TT	108 (35.9)	126 (39.6)	1 (referent)	—
		TC	154 (51.2)	154 (48.4)	1.17 (.83-1.64)	.89
		CC	39 (12.9)	38 (11.9)	1.20 (.72-2.00)	.68
-899T>G	rs762624	TT	160 (54.4)	168 (53.3)	1 (referent)	—
		TG	113 (38.4)	123 (39.0)	0.96 (.69-1.35)	.83
		GG	21 (7.1)	24 (7.6)	0.92 (.49-1.71)	.79
-791T>C	rs2395655	TT	94 (31.5)	112 (35.6)	1 (referent)	—
		TC	155 (52.0)	156 (49.5)	1.18 (.83-1.68)	.35
		CC	49 (16.5)	47 (14.9)	1.24 (.76-2.02)	.38
CDKN1B						
-1857C>T	rs3759217	CC	238 (79.6)	243 (77.1)	1 (referent)	—
		CT	58 (19.4)	71 (22.5)	0.83 (.56-1.23)	.36
		TT	3 (1.0)	1 (0.3)	3.06 (.32-29.65)	.33
-1608G>A	rs35756741	GG	245 (80.3)	267 (84.0)	1 (referent)	—
		GA	60 (19.7)	47 (14.8)	1.39 (.91-2.11)	.12
		AA	0 (0.0)	4 (1.2)	—	—
-373G>T	rs36228499	GG	93 (31.2)	99 (31.0)	1 (referent)	—
		GT	159 (53.4)	163 (51.1)	1.04 (.73-1.48)	.84
		TT	46 (15.4)	57 (17.9)	0.86 (.53-1.39)	.54
CDKN2A						
-222T>A	rs36228834	TT	266 (86.6)	298 (93.7)	1 (referent)	—
		TA	39 (12.7)	19 (6.0)	2.30 (1.30-4.08)	.004
		AA	2 (0.7)	1 (0.3)	2.24 (.20-24.85)	.51
CDKN2B						
-1270C>T	rs36229158	CC	277 (91.4)	302 (94.7)	1 (referent)	—
		CT	24 (7.9)	16 (5.0)	1.64 (.85-3.14)	.14
		TT	2 (0.7)	1 (0.3)	2.18 (.20-24.18)	.52

-593A>T,C	rs2069416	AA	129 (43.1)	107 (34.1)	1 (referent)	—
		AT	124 (41.5)	150 (47.8)	0.69 (.48-.97)	.034
		AC	11 (3.7)	12 (3.8)	0.60 (.35-1.03)	.07
		TT	29 (9.7)	40 (12.7)	0.76 (.32-1.79)	.53
		TC	6 (2.0)	5 (1.6)	0.99 (.30-3.35)	.99
		CC	0	0	—	—
-287G>C	rs2069418	GG	102 (34.1)	97 (30.7)	1 (referent)	—
		GC	141 (47.2)	168 (53.2)	0.80 (.56-1.14)	.22
		CC	56 (18.7)	51 (16.1)	1.04 (.65-1.67)	.86
E2F1						
-187C>T	rs3213141	CC	172 (57.5)	180 (57.0)	1 (referent)	—
		CT	119 (29.8)	120 (38.0)	1.04 (.75-1.44)	.82
		TT	8 (2.7)	16 (5.1)	0.52 (.22-1.25)	.15
HDAC1						
-1269T>C	rs1741981	TT	151 (49.8)	143 (44.1)	1 (referent)	—
		TC	133 (43.9)	148 (45.7)	0.85 (.61-1.18)	.33
		CC	19 (6.3)	33 (10.2)	0.54 (.30-1.00)	.051
-455T>C	rs36212119	TT	261 (83.4)	277 (84.2)	1 (referent)	—
		TC	49 (15.6)	49 (14.9)	1.06 (.69-1.63)	.79
		CC	3 (1.0)	3 (0.9)	1.06 (.21-5.30)	.94
MDM2						
-1494A>G	rs1144944	AA	73 (23.2)	82 (24.9)	1 (referent)	—
		AG	165 (52.4)	170 (51.7)	1.09 (.74-1.60)	.66
		GG	77 (24.4)	77 (23.4)	1.12 (.72-1.76)	.61
-1174DEL AAAAAGC(40bp)	rs3730485	AAAAAGC	104 (33.6)	119 (36.4)	1 (referent)	—
		AAAAAGC/-	156 (50.5)	151 (46.2)	1.18 (.84-1.67)	.34
		-/-	49 (15.9)	57 (17.4)	0.98 (.62-1.56)	.94
-182C>G	rs937282	CC	78 (24.9)	84 (25.6)	1 (referent)	—
		CG	164 (52.4)	171 (52.1)	1.03 (.71-1.50)	.87
		GG	71 (22.7)	73 (22.3)	1.05 (.67-1.64)	.84
+309T>G	rs2279744	TT	141 (45.1)	125 (38.5)	1 (referent)	—
		TG	139 (44.4)	162 (49.8)	0.76 (.55-1.06)	.11
		GG	33 (10.5)	38 (11.7)	0.77 (.45-1.30)	.33
SMAD3						
-1938T>C	rs36221701	TT	246 (78.8)	254 (77.2)	1 (referent)	—
		TC	62 (19.9)	66 (20.1)	0.97 (.66-1.43)	.88
		CC	4 (1.3)	9 (2.7)	0.46 (.14-1.51)	.20
RB1						
-1554C>A	rs1573601	CC	175 (58.1)	176 (55.3)	1 (referent)	—
		CA	112 (37.2)	126 (39.6)	0.89 (.64-1.24)	.51
		AA	14 (4.7)	16 (5.0)	0.88 (.42-1.86)	.74
TGFB1						
-1886A>G	rs2317130	AA	138 (44.2)	144 (50.0)	1 (referent)	—
		AG	135 (43.3)	151 (47.2)	0.93 (.67-1.30)	.68
		GG	39 (12.5)	25 (7.8)	1.63 (.94-2.83)	.09
-1571G>A	rs4803457	GG	105 (34.3)	113 (35.3)	1 (referent)	—
		GA	150 (49.0)	162 (50.6)	1.00 (.70-1.41)	.98
		AA	51 (16.7)	45 (14.1)	1.22 (.75-1.97)	.42

-1550DEL AGG	rs11466313	AGG/AGG	140 (45.4)	149 (46.1)	1 (referent)	—
		AGG/-	133 (43.2)	148 (45.8)	0.96 (.69-1.33)	.79
		-/-	35 (11.4)	26 (8.1)	1.43 (.82-2.50)	.21
-508G>A	rs1800469	GG	138 (44.4)	147 (45.5)	1 (referent)	—
		GA	136 (43.7)	150 (46.4)	.97 (.70-1.34)	.84
		AA	37 (11.9)	26 (8.1)	1.52 (.87-2.63)	.14
DNA double-strand break repair:						
ATM						
-1206G>T	rs4987876	GG	231 (79.9)	261 (82.3)	1 (referent)	—
		GT	54 (18.7)	51 (16.1)	1.20 (.78-1.82)	.29
		TT	4 (1.4)	5 (1.6)	0.90 (.24-3.41)	.88
-635T>A	rs228589	TT	111 (39.4)	111 (35.1)	1 (referent)	—
		TA	137 (48.6)	146 (46.2)	0.94 (.66-1.33)	.68
		AA	34 (12.0)	59 (18.7)	0.58 (.35-.95)	.03
BRCA1						
-1890T>C	rs4793204	TT	130 (42.6)	149 (46.4)	1 (referent)	—
		TC	144 (47.2)	136 (42.4)	1.21 (.87-1.69)	.25
		CC	31 (10.2)	36 (11.2)	0.99 (.58-1.68)	.96
-708A>G	rs799906	AA	124 (40.9)	145 (44.6)	1 (referent)	—
		AG	146 (48.2)	143 (44.0)	1.19 (.86-1.66)	.30
		GG	33 (10.9)	37 (11.4)	1.04 (.61-1.77)	.88
-598INS ACA	rs8176071	-/-	130 (42.5)	151 (46.2)	1 (referent)	—
		-/ACA	145 (47.4)	141 (43.1)	1.19 (.86-1.66)	.29
		ACA/ACA	31 (10.1)	35 (10.7)	1.03 (.60-1.76)	.92
-588A>G	rs3092986	AA	272 (89.5)	268 (82.7)	1 (referent)	—
		AG	31 (10.2)	54 (16.7)	0.56 (.35-.91)	.018
		GG	1 (0.3)	2 (0.6)	0.49 (.04-5.47)	.56
BRCA2						
-1555C>A	rs206114	CC	112 (35.5)	116 (35.3)	1 (referent)	—
		CA	133 (42.2)	154 (46.8)	0.89 (.63-1.27)	.53
		AA	70 (22.2)	59 (17.9)	1.23 (.80-1.89)	.35
-1260DEL GTCTAA	rs36221753	GTCTAA/GTCTAA	115 (36.7)	117 (35.8)	1 (referent)	—
		GTCTAA/-	137 (43.8)	161 (49.2)	.86 (.61-1.22)	.41
		-/-	61 (19.5)	49 (15.0)	1.27 (.80-2.00)	.31
-1144A>G	rs206115	AA	97 (32.5)	110 (34.5)	1 (referent)	—
		AG	134 (45.0)	157 (49.2)	.97 (.68-1.38)	.86
		GG	67 (22.5)	52 (16.3)	1.46 (.93-2.30)	.10
-1134C>T	rs206116	CC	105 (36.3)	111 (35.1)	1 (referent)	—
		CT	122 (42.2)	156 (49.4)	0.83 (.58-1.18)	.30
		TT	62 (21.5)	49 (15.5)	1.34 (.84-2.12)	.21
-908C>T	rs206117	CC	108 (35.5)	115 (35.8)	1 (referent)	—
		CT	135 (44.4)	157 (48.9)	0.91 (.64-1.30)	.62
		TT	61 (20.1)	49 (15.3)	1.32 (.84-2.10)	.23
-254G>A	rs3092989	GG	220 (69.4)	205 (62.3)	1 (referent)	—
		GA	89 (28.1)	113 (34.3)	0.73 (.52-1.03)	.07
		AA	8 (2.5)	11 (3.3)	0.68 (.27-1.72)	.41

RAD51						
-1185A>T	rs2619679	AA	76 (24.3)	81 (25.0)	1 (referent)	—
		AT	161 (51.4)	168 (51.9)	1.02 (.70-1.49)	.91
		TT	76 (24.3)	75 (23.1)	1.08 (.69-1.69)	.74
XRCC4						
-1864T>C	rs3763063	TT	98 (33.0)	79 (24.6)	1 (referent)	—
		TC	140 (47.1)	166 (51.7)	0.68 (.47-.98)	.042
		CC	59 (19.9)	76 (23.7)	0.63 (.40-.98)	.042
-1407C>G	rs1993947	CC	229 (78.7)	266 (83.9)	1 (referent)	—
		CG	59 (20.3)	49 (15.5)	1.40 (.92-2.12)	.12
		GG	3 (1.0)	2 (0.6)	1.74 (.29-10.52)	.54
XRCC5						
-1379G>T	rs828907	GG	101 (32.2)	88 (26.8)	1 (referent)	—
		GT	153 (48.7)	171 (52.1)	0.78 (.54-1.12)	.17
		TT	60 (19.1)	69 (21.1)	0.76 (.48-1.19)	.22
-297C>T	rs11685387	CC	140 (44.9)	178 (54.3)	1 (referent)	—
		CT	139 (44.5)	132 (40.2)	1.34 (.97-1.85)	.08
		TT	33 (10.6)	18 (5.5)	2.33 (1.26-4.31)	.007
XRCC6						
-1469C>T	rs28384701	CC	267 (86.7)	297 (92.2)	1 (referent)	—
		CT	41 (13.3)	21 (6.5)	2.17 (1.25-3.77)	.006
		TT	0 (0.0)	4 (1.2)	—	—
-1296C>G	rs2267437	CC	97 (31.5)	115 (35.9)	1 (referent)	—
		CG	155 (50.3)	163 (50.9)	1.13 (.80-1.60)	.50
		GG	56 (18.2)	42 (13.1)	1.58 (.98-2.56)	.06
-731G>A	rs132770	GG	197 (62.5)	189 (57.8)	1 (referent)	—
		GA	104 (33.0)	112 (34.2)	0.89 (.64-1.24)	.50
		AA	14 (4.5)	26 (8.0)	0.52 (.26-1.02)	.06

Percentages indicate number of individuals with a given genotype/total number of genotyped individuals. Nominally significant results are shown in bold.

Directionality of the association, as measured by the ORs, was used to define risk versus nonrisk genotype combinations.

OR indicates odds ratio; CI, confidence interval; —, not applicable.

Supplementary Table 2. Main effects of pSNPs on ALL risk among children, as estimated by hierarchical modeling with a range of second-stage residual variance values

Gene	High-risk genotype of tag pSNP	Case no. (n= 286)	Control no. (n= 297)	HM ORs (95% CIs) when second-stage variance equals:					
				$\tau^2 = 0.01$	$\tau^2 = 0.04$	$\tau^2 = 0.1225$	$\tau^2 = 0.25$	$\tau^2 = 0.4$	$\tau^2 = 1$
<i>CCND1</i>	-1938CC	67	60	1.21 (1.00-1.47)	1.23 (.92-1.66)	1.25 (.86-1.80)	1.25 (.84-1.87)	1.26 (.83-1.89)	1.26 (.82-1.93)
<i>CDC25A</i>	-2030GG	237	239	1.22 (1.00-1.49)	1.24 (.91-1.69)	1.24 (.84-1.83)	1.24 (.81-1.89)	1.24 (.79-1.91)	1.24 (.78-1.96)
<i>CDKN1A</i>	-1284TC/CC	181	178	1.21 (1.00-1.47)	1.23 (.93-1.63)	1.25 (.88-1.77)	1.26 (.86-1.84)	1.26 (.85-1.88)	1.27 (.84-1.92)
	-899TT	160	161	1.21 (1.00-1.46)	1.21 (.91-1.60)	1.21 (.86-1.71)	1.21 (.83-1.76)	1.21 (.82-1.78)	1.21 (.81-1.82)
<i>CDKN1B</i>	-1857CC	228	229	1.22 (1.00-1.48)	1.23 (.91-1.65)	1.22 (.84-1.78)	1.21 (.80-1.82)	1.20 (.78-1.84)	1.18 (.75-1.85)
	-1608GA/AA	56	50	1.19 (.98-1.45)	1.17 (.87-1.59)	1.15 (.78-1.70)	1.15 (.75-1.76)	1.14 (.73-1.79)	1.14 (.71-1.83)
	-373GG/GT	241	247	1.18 (.97-1.44)	1.14 (.84-1.56)	1.10 (.74-1.65)	1.08 (.69-1.69)	1.07 (.67-1.71)	1.05 (.64-1.73)
<i>CDKN2A</i>	-222TA/AA	40	19	1.27 (1.03-1.55)	1.40 (1.00-1.96)	1.58 (.99-2.51)	1.70 (1.00-2.89)	1.77 (1.01-3.12)	1.88 (1.02-3.45)
<i>CDKN2B</i>	-593AA/AC	136	113	1.27 (1.06-1.54)	1.38 (1.05-1.81)	1.45 (1.04-2.02)	1.47 (1.03-2.11)	1.48 (1.02-2.15)	1.49 (1.01-2.19)
	-287GG	96	94	1.21 (1.00-1.47)	1.23 (.93-1.63)	1.24 (.87-1.75)	1.23 (.85-1.80)	1.23 (.83-1.81)	1.22 (.81-1.83)
<i>E2F1</i>	-187CC	166	170	1.17 (.97-1.41)	1.13 (.86-1.47)	1.09 (.79-1.50)	1.07 (.76-1.50)	1.06 (.74-1.50)	1.04 (.73-1.49)
<i>HDAC1</i>	-1269TT	147	131	1.25 (1.03-1.50)	1.31 (1.00-1.71)	1.36 (.99-1.88)	1.39 (.98-1.95)	1.40 (.98-1.99)	1.42 (.98-2.04)
	-455TC/CC	54	48	1.20 (.98-1.46)	1.19 (.88-1.62)	1.18 (.80-1.75)	1.18 (.77-1.81)	1.18 (.75-1.84)	1.18 (.74-1.88)
<i>SMAD3</i>	-1938TT	224	226	1.19 (.98-1.44)	1.18 (.88-1.58)	1.17 (.82-1.68)	1.17 (.79-1.73)	1.17 (.78-1.75)	1.17 (.77-1.78)
<i>MDM2</i>	-1494AG/GG	224	226	1.15 (.95-1.39)	1.08 (.79-1.48)	1.03 (.67-1.59)	1.00 (.60-1.67)	0.99 (.57-1.71)	0.97 (.53-1.78)
	-1174AAAAAGC/--	191	190	1.14 (.95-1.38)	1.09 (.81-1.46)	1.06 (.71-1.58)	1.05 (.66-1.68)	1.05 (.64-1.74)	1.05 (.61-1.82)
	+309TT	127	111	1.22 (1.01-1.46)	1.26 (.96-1.65)	1.30 (.93-1.82)	1.32 (.92-1.90)	1.33 (.92-1.93)	1.34 (.91-1.97)
<i>RB1</i>	-1554CC	162	161	1.19 (.99-1.44)	1.18 (.90-1.54)	1.17 (.85-1.61)	1.16 (.83-1.63)	1.16 (.82-1.64)	1.16 (.81-1.65)
<i>TGFB1</i>	-1886GG	37	23	1.23 (1.01-1.51)	1.32 (.93-1.85)	1.48 (.89-2.46)	1.66 (.89-3.10)	1.80 (.89-3.64)	2.10 (.92-4.79)
	-1571AA	48	41	1.19 (.97-1.45)	1.15 (.83-1.59)	1.06 (.67-1.68)	0.97 (.56-1.69)	0.91 (.49-1.68)	0.81 (.40-1.64)
<i>ATM</i>	-1206GT/TT	52	52	1.22 (1.00-1.50)	1.15 (.84-1.56)	1.07 (.72-1.58)	1.03 (.67-1.58)	1.01 (.64-1.57)	0.98 (.61-1.57)
	-635TT	121	109	1.27 (1.05-1.54)	1.27 (.97-1.67)	1.29 (.93-1.78)	1.29 (.91-1.84)	1.30 (.91-1.87)	1.31 (.90-1.91)
<i>BRCA1</i>	-598-ACA/ACA ACA	164	158	1.25 (1.03-1.51)	1.21 (.93-1.59)	1.17 (.85-1.62)	1.15 (.82-1.62)	1.14 (.81-1.62)	1.13 (.79-1.63)
	-588AA	256	246	1.33 (1.08-1.64)	1.45 (1.05-2.00)	1.60 (1.05-2.43)	1.68 (1.06-2.68)	1.73 (1.06-2.81)	1.79 (1.07-2.99)
<i>BRCA2</i>	-1260--	59	47	1.29 (1.05-1.58)	1.33 (.97-1.82)	1.39 (.92-2.08)	1.42 (.90-2.24)	1.44 (.89-2.32)	1.46 (.88-2.42)
	-1144GG	63	51	1.28 (1.05-1.56)	1.29 (.95-1.76)	1.32 (.89-1.97)	1.34 (.86-2.09)	1.36 (.85-2.16)	1.37 (.84-2.24)
	-254GG	199	187	1.25 (1.04-1.51)	1.21 (.92-1.60)	1.17 (.82-1.66)	1.14 (.78-1.68)	1.13 (.76-1.69)	1.12 (.73-1.71)

<i>RAD51</i>	-1185TT	71	68	1.23 (1.01-1.50)	1.17 (.88-1.57)	1.13 (.79-1.62)	1.11 (.76-1.64)	1.11 (.74-1.65)	1.11 (.73-1.67)
<i>XRCC4</i>	-1864TT	102	76	1.33 (1.10-1.62)	1.42 (1.07-1.88)	1.49 (1.06-2.09)	1.51 (1.05-2.19)	1.52 (1.04-2.23)	1.53 (1.03-2.27)
	-1407CG/GG	67	50	1.29 (1.06-1.57)	1.30 (.96-1.76)	1.30 (.89-1.90)	1.30 (.86-1.96)	1.30 (.84-1.99)	1.30 (.83-2.03)
<i>XRCC5</i>	-1379GG	91	78	1.26 (1.04-1.52)	1.23 (.92-1.63)	1.20 (.84-1.71)	1.19 (.81-1.74)	1.18 (.79-1.76)	1.18 (.78-1.78)
	-297CT/TT	158	136	1.30 (1.07-1.56)	1.33 (1.02-1.74)	1.36 (.98-1.88)	1.37 (.97-1.95)	1.38 (.96-1.98)	1.38 (.95-2.01)
<i>XRCC6</i>	-1296GG	50	39	1.28 (1.04-1.56)	1.30 (.95-1.78)	1.34 (.88-2.03)	1.36 (.85-2.16)	1.37 (.84-2.23)	1.37 (.82-2.30)
	-731GG	178	170	1.25 (1.03-1.51)	1.22 (.93-1.60)	1.19 (.86-1.66)	1.18 (.83-1.69)	1.18 (.81-1.70)	1.17 (.80-1.72)
Pathway estimation									
DSBR				1.27 (1.14-1.42)	1.28 (1.10-1.47)	1.28 (1.04-1.58)	1.28 (.97-1.70)	1.28 (.91-1.82)	1.29 (.75-2.19)
Cell cycle				1.20 (1.09-1.33)	1.21 (1.07-1.38)	1.22 (1.02-1.46)	1.23 (.96-1.56)	1.23 (.92-1.65)	1.23 (.79-1.93)

Hierarchical modeling with pathway indicators only and τ^2 set to 0.01, 0.04, 0.1225, 0.25, 0.4 or 1 which corresponds to being 95% certain a priori that an OR lies in a 1.5-fold, 2.2-fold, 4-fold, 7-fold, a 12-fold and 50-fold range, respectively (note that $\exp(3.92\tau) = 2.2$ if $\tau^2 = 0.04$). Significant results are shown in bold.

HM, hierarchical model; OR, odds ratio; CI, confidence interval; DSBR, double strand break repair.

Supplementary Table 3. Gene-gene interaction effects of selected polymorphisms on ALL risk among children, as estimated by hierarchical modeling with a range of second-stage residual variance

Gene, gene-gene interactions	Case no. (n= 286)	Control no. (n= 297)	HM ORs (95% CIs) when second-stage variance equals:					
			$\tau^2 = 0.01$	$\tau^2 = 0.04$	$\tau^2 = 0.1225$	$\tau^2 = 0.25$	$\tau^2 = 0.4$	$\tau^2 = 1$
<i>CDKN2A-222TA/AA</i>	40	19	1.10 (.89-1.35)	1.18 (.82-1.69)	1.32 (.74-2.36)	1.50 (.68-3.28)	1.67 (.64-4.35)	2.19 (.55-8.68)
<i>CDKN2B-593AA/AC</i>	136	113	1.07 (.88-1.29)	1.10 (.80-1.51)	1.20 (.74-1.95)	1.33 (.72-2.47)	1.46 (.72-2.96)	1.81 (.76-4.34)
<i>HDAC1-1269TT</i>	147	131	1.04 (.86-1.26)	1.03 (.75-1.41)	1.03 (.63-1.67)	1.03 (.55-1.91)	1.03 (.51-2.10)	1.06 (.45-2.53)
<i>BRCA1-588AA</i>	256	246	1.14 (.93-1.40)	1.18 (.85-1.65)	1.23 (.76-1.98)	1.28 (.71-2.32)	1.35 (.69-2.65)	1.57 (.68-3.65)
<i>XRCC4-1864TT</i>	102	76	1.13 (.93-1.36)	1.17 (.85-1.62)	1.29 (.78-2.14)	1.45 (.76-2.77)	1.61 (.76-3.40)	2.05 (.80-5.21)
<i>XRCC5-297CT/TT</i>	158	136	1.12 (.93-1.35)	1.15 (.84-1.58)	1.24 (.77-2.00)	1.36 (.74-2.49)	1.47 (.73-2.95)	1.76 (.74-4.16)
<i>CDKN2A-222TA/AA*CDKN2B-593AA/AC</i>	25	12	1.07 (.86-1.32)	1.07 (.74-1.55)	1.03 (.58-1.83)	0.96 (.46-2.01)	0.91 (.39-2.10)	0.78 (.28-2.16)
<i>CDKN2A-222TA/AA*HDAC1-1269TT</i>	15	7	1.06 (.85-1.31)	1.05 (.72-1.54)	1.01 (.55-1.84)	0.96 (.45-2.06)	0.92 (.39-2.18)	0.84 (.30-2.37)
<i>CDKN2A-222TA/AA*BRCA1-588AA</i>	38	18	1.22 (1.00-1.49)	1.29 (.90-1.85)	1.40 (.78-2.49)	1.48 (.68-3.24)	1.55 (.60-3.96)	1.61 (.43-6.04)
<i>CDKN2A-222TA/AA*XRCC4-1864TT</i>	16	4	1.20 (.98-1.48)	1.26 (.86-1.85)	1.35 (.73-2.49)	1.45 (.66-3.19)	1.52 (.61-3.78)	1.67 (.54-5.13)
<i>CDKN2A-222TA/AA*XRCC5-297CT/TT</i>	17	12	1.15 (.94-1.42)	1.06 (.73-1.54)	0.87 (.48-1.56)	0.70 (.33-1.46)	0.59 (.25-1.37)	0.43 (.16-1.17)
<i>CDKN2B-593AA/AC*HDAC1-1269TT</i>	69	51	1.05 (.86-1.27)	1.04 (.75-1.44)	1.04 (.65-1.64)	1.03 (.60-1.77)	1.03 (.58-1.85)	1.02 (.54-1.94)
<i>CDKN2B-593AA/AC*BRCA1-588AA</i>	120	93	1.18 (.98-1.42)	1.17 (.85-1.61)	1.12 (.69-1.81)	1.05 (.57-1.92)	0.99 (.50-1.96)	0.85 (.37-1.95)
<i>CDKN2B-593AA/AC*XRCC4-1864TT</i>	45	30	1.15 (.95-1.40)	1.09 (.77-1.53)	0.98 (.60-1.61)	0.90 (.50-1.62)	0.85 (.45-1.60)	0.76 (.38-1.54)
<i>CDKN2B-593AA/AC*XRCC5-297CT/TT</i>	70	48	1.16 (.99-1.41)	1.13 (.81-1.56)	1.08 (.68-1.72)	1.05 (.61-1.80)	1.03 (.58-1.85)	0.99 (.52-1.89)
<i>HDAC1-1269TT*BRCA1-588AA</i>	133	106	1.19 (.99-1.43)	1.23 (.90-1.69)	1.31 (.81-2.10)	1.37 (.76-2.50)	1.42 (.72-2.80)	1.49 (.66-3.38)
<i>HDAC1-1269TT*XRCC4-1864TT</i>	56	38	1.16 (.96-1.41)	1.13 (.80-1.58)	1.07 (.66-1.74)	1.02 (.58-1.81)	0.99 (.53-1.84)	0.94 (.47-1.86)
<i>HDAC1-1269TT*XRCC5-297CT/TT</i>	81	62	1.14 (.95-1.38)	1.09 (.79-1.50)	1.00 (.64-1.58)	0.94 (.56-1.61)	0.91 (.51-1.61)	0.85 (.45-1.60)
<i>BRCA1-588AA*XRCC4-1864TT</i>	90	64	1.12 (.92-1.35)	1.11 (.80-1.54)	1.08 (.65-1.77)	1.02 (.54-1.91)	0.96 (.47-1.98)	0.83 (.34-1.99)
<i>BRCA1-588AA*XRCC5-297CT/TT</i>	141	113	1.12 (.94-1.35)	1.14 (.84-1.55)	1.14 (.72-1.83)	1.12 (.62-2.03)	1.09 (.56-2.14)	1.00 (.44-2.28)
<i>XRCC4-1864TT *XRCC5-297CT/TT</i>	55	35	1.11 (.91-1.35)	1.10 (.79-1.54)	1.09 (.67-1.76)	1.06 (.60-1.88)	1.04 (.56-1.93)	0.98 (.50-1.95)

Hierarchical modeling of significant main effects plus pair-wise gene-gene interaction effects using pathway indicators only and with τ^2 set to 0.01, 0.04, 0.1225, 0.25, 0.4 or 1 which corresponds to being 95% certain a priori that an OR lies in a 1.5-fold, 2.2-fold, 4-fold, 7-fold, 12-fold and 50-fold range, respectively (note that $\exp(3.92\tau) = 4$ if $\tau^2 = 0.1225$).

Significant results are shown in bold.

HM, hierarchical model; OR, odds ratio; CI, confidence interval

Supplementary Table 4. Gene-gene interaction effects on pre-B ALL risk among children, as estimated by conventional maximum-likelihood analysis and hierarchical modeling

Gene, high-risk genotype gene-gene interaction	Case no. (n= 286)	Control no. (n= 297)	Conventional GxG 1, OR (95% CI)	Conventional GxG 2, OR (95% CI)	Backward stepwise, OR (95% CI)	Pathway HM 1, OR (95% CI)
<i>CDKN2A-222TA/AA</i>	40	19	2.38 (1.34-4.22)	6.92 (.33-144.02)	4.39 (1.79-10.73)	1.32 (.74-2.36)
<i>CDKN2B-593AA/AC</i>	136	113	1.48 (1.06-2.05)	3.02 (.94-9.74)	1.46 (1.03-2.05)	1.20 (.74-1.95)
<i>HDAC1-1269TT</i>	147	131	1.34 (.97-1.86)	1.15 (.38-3.49)	1.35 (.96-1.89)	1.03 (.63-1.67)
<i>BRCA1-588AA</i>	256	246	1.77 (1.09-2.87)	2.54 (.79-8.22)	1.72 (1.05-2.83)	1.23 (.76-1.98)
<i>XRCC4-1864TT</i>	102	76	1.61 (1.13-2.30)	3.53 (1.02-12.20)	1.70 (1.20-2.42)	1.29 (.78-2.14)
<i>XRCC5-297CT/TT</i>	158	136	1.46 (1.05-2.02)	2.75 (.89-8.51)	1.70 (1.20-2.42)	1.24 (.77-2.00)
<i>CDKN2A-222TA/AA * CDKN2B-593AA/AC</i>	25	12	0.67 (.21-2.20)	0.55 (.16-1.96)	–	1.03 (.58-1.83)
<i>CDKN2A-222TA/AA * HDAC1-1269TT</i>	15	7	0.72 (.22-2.33)	0.68 (.19-2.48)	–	1.01 (.55-1.84)
<i>CDKN2A-222TA/AA * BRCA1-588AA</i>	38	18	0.62 (.05-7.65)	0.90 (.05-15.81)	–	1.40 (.78-2.49)
<i>CDKN2A-222TA/AA * XRCC4-1864TT</i>	16	4	1.63 (.43-6.12)	1.82 (.44-7.57)	–	1.35 (.73-2.49)
<i>CDKN2A-222TA/AA * XRCC5-297CT/TT</i>	17	12	0.26 (.08-.84)	0.26 (.08-.89)	0.26 (.08-.85)	0.87 (.48-1.56)
<i>CDKN2B-593AA/AC * HDAC1-1269TT</i>	69	51	0.89 (.46-1.72)	0.97 (.48-1.96)	–	1.04 (.65-1.64)
<i>CDKN2B-593AA/AC * BRCA1-588AA</i>	120	93	0.82 (.31-2.18)	0.58 (.20-1.70)	–	1.12 (.69-1.81)
<i>CDKN2B-593AA/AC * XRCC4-1864TT</i>	45	30	0.74 (.36-1.53)	0.62 (.29-1.35)	–	0.98 (.60-1.61)
<i>CDKN2B-593AA/AC * XRCC5-297CT/TT</i>	70	48	0.93 (.48-1.80)	0.92 (.45-1.85)	–	1.08 (.68-1.72)
<i>HDAC1-1269TT * BRCA1-588AA</i>	133	106	1.57 (.60-4.14)	1.55 (.55-4.33)	–	1.31 (.81-2.10)
<i>HDAC1-1269TT * XRCC4-1864TT</i>	56	38	0.90 (.44-1.85)	0.85 (.40-1.79)	–	1.07 (.66-1.74)
<i>HDAC1-1269TT * XRCC5-297CT/TT</i>	81	62	0.88 (.46-1.70)	0.77 (.39-1.53)	–	1.00 (.64-1.58)
<i>BRCA1-588AA * XRCC4-1864TT</i>	90	64	0.71 (.25-2.03)	0.57 (.18-1.79)	–	1.08 (.65-1.77)
<i>BRCA1-588AA * XRCC5-297CT/TT</i>	141	113	0.91 (.34-2.40)	0.75 (.26-2.15)	–	1.14 (.72-1.83)
<i>XRCC4-1864TT * XRCC5-297CT/TT</i>	55	35	0.91 (.44-1.85)	0.86 (.41-1.81)	–	1.09 (.67-1.76)

Conventional GxG model 1, logistic regression with a single marker in each regression model or both main effects and the product interaction term in which case interaction ORs (95% CIs) are shown; Conventional GxG model 2, logistic

regression with all markers and product interaction terms in one single model; Backward stepwise, backward stepwise logistic regression forcing all significant main effects into the model and using $P \leq 0.05$ for removal from and $P \leq 0.10$ for addition to the model; Pathway HM 1; hierarchical modeling with τ^2 set to 0.1225 and pathway indicators only. Significant results are shown in bold.

GxG, gene-gene interaction; HM, hierarchical modeling; OR, odds ratio; CI, confidence interval.

References

1. Krajinovic M, Labuda D, Richer C, Karimi S, Sinnett D. Susceptibility to childhood acute lymphoblastic leukemia: influence of CYP1A1, CYP2D6, GSTM1, and GSTT1 genetic polymorphisms. *Blood* 1999;93:1496-501.
2. Krajinovic M, Sinnett H, Richer C, Labuda D, Sinnett D. Role of NQO1, MPO and CYP2E1 genetic polymorphisms in the susceptibility to childhood acute lymphoblastic leukemia. *Int J Cancer* 2002;97:230-6.
3. Batar B, Guven M, Baris S, Celkan T, Yildiz I. DNA repair gene XPD and XRCC1 polymorphisms and the risk of childhood acute lymphoblastic leukemia. *Leuk Res* 2009;33:759-63.
4. Petra BG, Janez J, Vita D. Gene-gene interactions in the folate metabolic pathway influence the risk for acute lymphoblastic leukemia in children. *Leuk Lymphoma* 2007;48:786-92.
5. Healy J, Belanger H, Beaulieu P, Lariviere M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood* 2007;109:683-92.
6. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet* 2009;41:1006-10.
7. Trevino LR, Yang W, French D, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet* 2009;41:1001-5.
8. Sherborne AL, Hosking FJ, Prasad RB, et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. *Nat Genet* 2010.
9. Han S, Lee KM, Park SK, et al. Genome-wide association study of childhood acute lymphoblastic leukemia in Korea. *Leuk Res* 2010.
10. Malumbres M, Barbacid M. To cycle or not to cycle: a critical decision in cancer. *Nat Rev Cancer* 2001;1:222-31.
11. Hoeijmakers JH. Genome maintenance mechanisms for preventing cancer. *Nature* 2001;411:366-74.
12. Papaefthymiou MA, Giaginis CT, Theocharis SE. DNA repair alterations in common pediatric malignancies. *Med Sci Monit* 2008;14:RA8-15.

13. Okuda T, Shurtleff SA, Valentine MB, et al. Frequent deletion of p16INK4a/MTS1 and p15INK4b/MTS2 in pediatric acute lymphoblastic leukemia. *Blood* 1995;85:2321-30.
14. Greaves MF, Wiemels J. Origins of chromosome translocations in childhood leukaemia. *Nat Rev Cancer* 2003;3:639-49.
15. Greenland S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. *Stat Med* 1992;11:219-30.
16. Witte JS, Greenland S, Haile RW, Bird CL. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology* 1994;5:612-21.
17. Hung RJ, Brennan P, Malaveille C, et al. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiol Biomarkers Prev* 2004;13:1013-21.
18. Healy J, Richer C, Bourgey M, Kritikou EA, Sinnett D. Replication analysis confirms the association of ARID5B with childhood B-cell acute lymphoblastic leukemia. *Haematologica* 2010.
19. Baccichet A, Qualman SK, Sinnett D. Allelic loss in childhood acute lymphoblastic leukemia. *Leuk Res* 1997;21:817-23.
20. Rockman MV, Wray GA. Abundant raw material for cis-regulatory evolution in humans. *Mol Biol Evol* 2002;19:1991-2004.
21. Koo SH, Ong TC, Chong KT, Lee CG, Chew FT, Lee EJ. Multiplexed genotyping of ABC transporter polymorphisms with the Bioplex suspension array. *Biol Proced Online* 2007;9:27-42.
22. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* 1998;63:259-66.
23. Becker T, Knapp M. Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet Epidemiol* 2004;27:21-32.
24. Witte JS, Greenland S, Kim LL. Software for hierarchical modeling of epidemiologic data. *Epidemiology* 1998;9:563-6.
25. Witte JS, Greenland S, Kim LL, Arab L. Multilevel modeling in epidemiology with GLIMMIX. *Epidemiology* 2000;11:684-8.

26. Levine M, Davidson EH. Gene regulatory networks for development. *Proc Natl Acad Sci U S A* 2005;102:4936-42.
27. Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res* 2010;20:565-77.
28. Blanchette M, Bataille AR, Chen X, et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* 2006;16:656-68.
29. Ferretti V, Poitras C, Bergeron D, Coulombe B, Robert F, Blanchette M. PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res* 2007;35:D122-6.
30. Belanger H, Beaulieu P, Moreau C, Labuda D, Hudson TJ, Sinnett D. Functional promoter SNPs in cell cycle checkpoint genes. *Hum Mol Genet* 2005;14:2641-8.
31. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009;461:747-53.
32. Huebinger RM, Garner HR, Barber RC. Pathway genetic load allows simultaneous evaluation of multiple genetic associations. *Burns* 2010;doi:10.1016/j.burns.2010.02.001.
33. Mullighan CG, Downing JR. Genome-wide profiling of genetic alterations in acute lymphoblastic leukemia: recent insights and future directions. *Leukemia* 2009;23:1209-18.
34. Wada T, Kikuchi J, Nishimura N, Shimizu R, Kitamura T, Furukawa Y. Expression levels of histone deacetylases determine the cell fate of hematopoietic progenitors. *J Biol Chem* 2009;284:30673-83.
35. Huen MS, Sy SM, Chen J. BRCA1 and its toolbox for the maintenance of genome integrity. *Nat Rev Mol Cell Biol* 2010;11:138-48.
36. Friedenson B. The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. *BMC Cancer* 2007;7:152.
37. Deutsch E, Jarrousse S, Buet D, et al. Down-regulation of BRCA1 in BCR-ABL-expressing hematopoietic cells. *Blood* 2003;101:4583-8.
38. Rassool FV. DNA double strand breaks (DSB) and non-homologous end joining (NHEJ) pathways in human leukemia. *Cancer Lett* 2003;193:1-9.

39. Wu KH, Wang CH, Yang YL, et al. Significant association of XRCC4 single nucleotide polymorphisms with childhood leukemia in Taiwan. *Anticancer Res* 2010;30:529-33.
40. Jones S, Zhang X, Parsons DW, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 2008;321:1801-6.

Reprint Permissions

This research has been submitted to *Cancer Research* and is pending review.
The manuscript is included with the permission of the authors.

CHAPTER FOUR

*THE ROLE OF MATERNALLY-MEDIATED
GENETIC EFFECTS IN THE
SUSCEPTIBILITY TO CHILDHOOD ACUTE
LYMPHOBLASTIC LEUKEMIA*

Another complex biologic mechanism that is important to consider in the study of diseases that originate during fetal life is the possible role that the mother's, and to a lesser extent the father's, genetics could play in disease development. Exposure and genetic make-up of the father could lead to increased mutational burden at the level of the gametes, affecting offspring risk at a prezygotic stage. More importantly perhaps, the mother plays a crucial role as not only the genetic donor but also as fetal environment. Thus a susceptibility allele acting through the mother's genotype could have an effect on the fetus by shaping the intrauterine milieu, regardless of whether the allele was passed on to the fetus. Discriminating between maternal genotype effects and inherited genetic effects acting through the child or from joint fetomaternal effects, is important to better understand the underlying disease mechanisms, however this is complicated by the inherent fact that case genotype effects can alias maternal genotype effects due to simple Mendelian inheritance.

In this third objective, I built upon existing methods of analysis to explore maternally-mediated genetic effects and fetomaternal relationships in childhood ALL susceptibility. Simulation studies were performed in collaboration with colleagues in the laboratory, to show how existing methods can be adapted to mixed data sets composed of unrelated cases and controls, as well as case-parent triads. We also showed how these methods perform when the assumption of mating symmetry (i.e. equal allele frequencies between mothers and fathers mating in the population, an assumption that is required to bar from spurious maternal genotype associations), cannot be directly verified. Finally I evaluated the role of maternal genetic effects in the susceptibility to childhood ALL and showed, for the first time, that the mother's genotype can influence the risk of leukemia among her offspring, further corroborating the importance of parental genetic contributions to the susceptibility to early-onset disorders such as childhood leukemia. This work lead to the publication of a manuscript which is presented in this chapter:

- Detection of fetomaternal genotype associations in early-onset disorders: evaluation of different methods and their application to childhood leukemia (Healy et al. (2010) *Journal of Biomedicine and Biotechnology*, doi:10.1155/2010/369534).

Tables and figures presenting the results of the fetomaternal association studies for the DNA double-strand break repair genes are available in Appendix III.

Detection of Fetomaternal Genotype Associations in Early-Onset Disorders: Evaluation of Different Methods and their Application to Childhood Leukemia

Jasmine Healy,^{1*} Mathieu Bourgey,^{1*} Chantal Richer,¹ Daniel Sinnett,^{1,2}
and Marie-Helene Roy-Gagnon^{1,3}

¹Division of Hematology-Oncology, Sainte-Justine Hospital Research Center, University of Montreal, QC, Canada;

²Department of Pediatrics, Faculty of Medicine, University of Montreal, QC, Canada

³Department of Social and Preventive Medicine, Faculty of Medicine, University of Montreal, QC, Canada;

* Jasmine Healy and Mathieu Bourgey contributed equally to this work

This was an invited article in the *Journal of Biomedicine and Biotechnology's* Special Issue on Gene-Gene Interactions in Maternal and Perinatal Research. Healy J, Bourgey M, Richer C, Sinnett D, Roy-Gagnon MH. Detection of fetomaternal genotype associations in early-onset disorders: evaluation of different methods and their application to childhood leukemia. J Biomed Biotech. 2010; doi:10.1155/2010/369534). © the Hindawi Publishing Corporation. The article is included with the permission of the authors and the journal.

Author Contributions

For this article, I carried out 70% of the work. I participated in the design of the project as well as the analysis and interpretation of the results. I performed some of the genotyping and also conducted the association study and performed the relevant statistical analyses. I wrote the manuscript. M. Bourgey and MH. Roy-Gagnon developed the statistic. M. Bourgey performed the simulation study and participated in the interpretation of the results. C. Richer performed some of the genotyping. MH. Roy-Gagnon and D. Sinnett contributed to the conception and design of the study. MH. Roy-Gagnon also contributed to the analysis and interpretation of results and is principal investigator.

Abstract

Several designs and analytical approaches have been proposed to dissect offspring from maternal genetic contributions to early-onset diseases. However, lack of parental controls halts the direct verification of the assumption of mating symmetry (MS) required to assess maternally-mediated effects. In this study, we used simulations to investigate the performance of existing methods under mating asymmetry (MA) when parents of controls are missing. Our results show that the log-linear, likelihood-based framework using a case-triad/case-control hybrid design provides valid tests for maternal genetic effects even under MA. Using this approach, we examined fetomaternal associations between 29 SNPs in 12 cell cycle genes and childhood pre-B acute lymphoblastic leukemia (ALL). We identified putative fetomaternal effects at loci *CDKN2A* rs36228834 ($P = .017$) and *CDKN2B* rs36229158 ($P = .022$) that modulate the risk of childhood ALL. These data further corroborate the importance of the mother's genotype on the susceptibility to early-onset diseases.

Introduction

The risk for early-onset disorders can be influenced both by the inherited genotype of the child as well as by parentally-mediated genetic effects (1). The mother has a crucial role in early-onset disease predisposition as she provides the prenatal environment (2, 3) and can influence her offspring's risk of disease not only as a genetic donor but also through the effects of her genes acting directly on the intrauterine milieu or indirectly through fetomaternal gene-gene interactions (1, 4-6). Given the important role the mother can play in shaping disease susceptibility in her offspring, focusing solely on the genotype of the child in association testing could, in certain instances, be misleading.

Several family-based tests have been proposed to dissect offspring and maternal genetic contributions to early-onset disorders including the case-parent designs of Wilcox et al. (1) and Weinberg et al. (3) using a log-linear framework, and of Cordell et al. (7, 8), which uses a conditional logistic regression framework. Designs using alternative family structures have also been suggested: the hybrid design based on augmenting a set of case-parent trios with a set of parents of unrelated controls (9), the "pent" design consisting of the affected child, mother, father, and maternal grandparents (2), the case-mother/control-mother dyad design (10) or the design consisting of case-parent triads supplemented by control-mother dyads (11). However, the use of these alternative designs can be limited by the difficulty of obtaining grandparental data or sufficiently large samples of parents of unaffected children, and by the increase in costs incurred for genotyping these additional sets of individuals. On the other hand, in many ongoing genetic association studies it often occurs that parents of cases and unrelated control individuals are collected in parallel, in which case designs based solely on case-triad data could lead to considerable losses in power if unrelated case-control data are disregarded. In the event that

both case-parent triads as well as case-control data are ascertained, additional designs/analytical strategies are needed for fetomaternal genotype association testing.

The first aim of this study was to evaluate the adaptability of existing methods to deal with mixed data sets consisting of both case-parent triads and case-control data. We used simulations to investigate the validity and power of 1) Weinberg and Umbach's hybrid design (9) treating parents of controls as missing (HD-NPC) and 2) a classic case-control test in conjunction with Cordell et al.'s conditional logistic regression method (7) (CC+CLR), to distinguish between offspring and maternal genetic contributions to disease. Given that Cordell et al.'s approach relies solely on the use of case-parent triads, combining it with a case-control test will allow us to maximize the use of available genotype information. Finally, we compared both these approaches to a third hypothetical, ideal situation, in which genotype data from parents of controls would be available and Weinberg and Umbach's hybrid design could be used (HD) as described in (9).

However lack of parents of controls in HD-NPC and CC+CLR precludes the direct verification of the assumption of mating symmetry (MS) that is required to assess maternally-mediated effects. Mating symmetry (MS) refers to the hypothesis that for a parental genotype pair, the frequency in the population for a given mother-father genotype assignment is the same as for the reverse father-mother assignment (1, 12). Departures from this symmetry could lead to genotype frequency differences among males and females mating in the population which, in the context of fetomaternal association testing, could lead to confounding and spurious maternal associations. As such, the HD approach, using an auxiliary sample of parents of controls to obtain direct information on mating frequencies (9), is the only method that allows for the assumption of

symmetry to be directly tested and readily accommodated. It is unclear however how this method performs when parents of controls are missing and MA is present. And since parents of unaffected controls are not available in most ongoing studies, it is important to assess the robustness of these fetomaternal association tests under such circumstances. Therefore, we evaluated type I error rates and power of the three methods (HD, HD-NPC and CC+CLR) under varying degrees of MA, and genotypic risk models involving child, mother or both child and mother jointly, in order to identify the analytical approach that is most reliable for dissecting child and maternal genetic contributions to early-onset diseases in the absence of parents of controls.

Another aim of this study was to use these methods to investigate fetomaternal associations in a real mixed dataset of childhood pre-B acute lymphoblastic leukemia (ALL) patients. ALL is a hematological malignancy resulting from chromosomal alterations and mutations that affect molecular pathways that disrupt lymphoid progenitor cell differentiation (13, 14). There is well-established evidence for prenatal initiation of the leukemogenesis process in children (15-18). Moreover, parental exposures to environmental carcinogens or use of medication have been identified as potential risk factors for childhood leukemia (19-23) and transplacental carcinogen exposure has been involved in the development of certain subtypes of ALL (24). Although the risk of leukemia from environmental exposures *in utero* or in early childhood is likely to be influenced by genetic variation at both the level of the child and the mother, the role of maternally-mediated genetic effects in childhood leukemia susceptibility remains undefined. Here we performed a candidate gene association study using both ALL case-parent triads and unrelated controls to assess the impact of 29 SNPs from 12 cell cycle genes in both mother and child on childhood pre-B ALL risk.

Materials and Methods

Methods used to test for fetomaternal genotype associations

We compared three analytical approaches for the detection of early-onset disease associations. In the event that both case-parent triads and unrelated case-control data are available, we tested 1) a combined method in which a case-control genotypic test was carried out in conjunction with the conditional logistic regression test of Cordell et al. to detect associations at the level of the child and mother, respectively (CC+CLR); and 2) the log-linear, likelihood-based approach of Weinberg and Umbach (9) using an additional set of unrelated cases and unrelated controls as proxies for parental control genotype information (HD-NPC). We also compared these two approaches to 3) one in which parents of controls are also available and therefore the hybrid design (case-parent/parents of controls) can be used through log-linear, likelihood-based analysis (HD).

It should be noted that the combined CC+CLR approach is not a modification of the conditional logistic regression approach of Cordell et al. but rather an adaptation in its use to detect fetomaternal associations. Cordell et al.'s approach relies solely on the use of case-parent triads. Since disregarding any available unrelated case-control genotype data reduces power, we used Pearson's chi-square tests or Fisher's exact tests (CC), as appropriate, and conditional logistic regression (CLR) in parallel on partially overlapping data. The former were used on all available cases to identify genotype associations at the level of the child whereas the latter was used on case-parent triads to identify maternally-contributed effects. Results from the two tests were not combined, rather if a significant association was found in the child CC test then the CLR test was used (albeit on a reduced case set) in order to dissect

offspring and maternal effects. Similarly, if a significant result was found for the mother test using CLR, then CLR was further used to distinguish a main effect of the mother from a joint fetomaternal effect (see likelihood-ratio testing below).

Likelihood-ratio testing to dissect child from maternal genetic effects

Given that the offspring will be enriched for the risk allele by simple Mendelian inheritance (25), it is important to discriminate between direct effects of a maternal genotype or of a child genotype from a joint fetomaternal effect. To do so, we used a forward stepwise likelihood-ratio testing procedure. In the first step, we performed two single-step tests to investigate associations at the level of the child and mother separately (Table 1). For CC+CLR, Pearson's chi-square or Fisher's exact tests were performed in R (version 2.6.2), to compare genotype distributions in cases (unrelated and triad cases) vs. controls. In parallel, we used case-parent triads and logistic regression conditioning on exchangeable parental genotypes (CEPG) (7, 8), using the DGCgenetics package for R developed by D. Clayton (available at http://www-gene.cimr.cam.ac.uk/clayton/software/DGCgenetics_1.0.tar.gz) to test for maternally-mediated associations through a two degree-of-freedom likelihood-ratio test (Mother vs. Null). For HD-NPC and HD, log-linear regression analysis was performed using the LEM software (Log-linear and event history analysis with missing data using the EM algorithm) (26) and LEM script files provided by Weinberg et al. (available at <http://www.niehs.nih.gov/research/resources/software/hybrid/index.cfm> (9, 26)). Two degree-of-freedom likelihood-ratio tests were performed to independently test for offspring (Child vs. Null) and maternal (Mother vs. Null) associations in the log-linear framework.

In the second step, if a significant effect was found in the first step (based on a Bonferroni correction of $P < 0.025$ for the 2 tests performed), then the most significant model was tested against a joint effects model (Child+Mother vs. Child or Child+Mother vs. Mother) in a two degree-of-freedom likelihood-ratio test (Table 1). For CC+CLR this meant that conditional logistic regression was used on the reduced case-triad data set to test either Child+Mother vs. Child or Child+Mother vs. Mother depending on the results from step one. Whereas log-linear regression analysis was performed on case-triads and unrelated cases and controls in HD-NPC, and on case-triads, unrelated cases and parents of controls in HD.

Simulated data

We simulated cohort data under several conditions to model different genotypic risk effects involving either child, mother or both child and mother (joint fetomaternal effects) (Table 2). Under each model 200 replicate datasets were simulated using the R software. Replicates that did not fit the simulated models were not included in the analysis. To imitate our childhood leukemia cohort, we simulated genotypic data for 200 case-parent triads, 130 unrelated cases and 325 unrelated control-parent triads for each replicate. For all of our calculations we used the same overall number of affected individuals however the methods differed in how the case and control data were utilized in each individual test (Figure 1). For all of our child-based tests we used the same number of affected individuals and unrelated controls (Figure 1): 330 cases (200 case-triads and 130 unrelated cases) were compared to 325 unrelated controls. However the number of cases and population-based controls used for maternal and fetomaternal association testing varied depending on the method: CC+CLR used the genotypes from 200 case-triads only; HD-NPC used the genotypes of 330 cases (200 case-triads and 130 unrelated cases) and of 325 unrelated

controls; and HD incorporated the genotypic information of the parents ($n=650$) rather than the unrelated controls themselves (Figure 1).

In HD and HD-NPC, unrelated cases and controls were used by treating their parental genotypes as missing and the EM algorithm implemented in the LEM software was used to infer missing genotype information (26). Unrelated cases and cases belonging to triads were considered to have similar penetrance and thus similar genotypic relative risks (GRR). We assumed Hardy-Weinberg equilibrium and the absence of population stratification in the form of admixture. We evaluated the different approaches in terms of type I error rate and power to detect associations by counting the number of replicates found to be significantly associated over the total number of replicates that fit the specified model.

Simulated scenarios

We first assessed the behaviour of the three methods in scenarios in which MS was assumed across parents. Under MS, HD-NPC and HD were performed by forcing six mating-type variables in the log-linear model (9).

A second set of simulations was performed in which we assessed the performance of the tests assuming differences in genotype frequencies between males and females mating in the population, i.e. assuming varying levels of mating asymmetry. MA was evaluated in terms of the degree of departure from the expected mate-pair probability under symmetry. We used C_i to denote the disequilibrium for the i^{th} parental mating type. C_i is a multiplicative factor between 0 and 2 that describes the over representation ($C > 1$), under representation ($C < 1$) or symmetry ($C = 1$) of a mate-pair combination in the corresponding i^{th} mating type. The level of departure from MS is denoted as ΔC ,

a numerical value ranging from -1 to 1 , with $C = 1 + \Delta C$. The expected parental genotype distributions under the assumption of MS and MA are shown in Table 3. By varying the departure from mating symmetry, ΔC , we introduced varying levels of asymmetry into our simulations (Table 2). In this study we assumed a model for MA where $C_1 = C_2$ and $C_4 = 1$. Supplementary Figure 1 shows how departures from symmetry, as measured by ΔC , translate into differences in reciprocal mating types and overall genotype frequencies between males and females mating in the population. Under MA, HD-NPC and HD were performed by forcing nine mating-type variables in the log-linear model (9) and we tested for the presence of asymmetry by comparing the nine mating-type model to the six mating-type model with a three degree-of-freedom likelihood-ratio test using a cut-off of $P < 0.05$ to declare deviation from symmetry. A less conservative P value < 0.10 was also used (data not shown) and yielded similar results.

Childhood acute lymphoblastic leukemia data

Study subjects. We investigated fetomaternal associations in a pre-B acute lymphoblastic leukemia (ALL) cohort. The study population has been previously described (27, 28). Briefly, incident cases of childhood pre-B ALL ($n=321$) were diagnosed in the Division of Hematology-Oncology of the Sainte-Justine Hospital in Montreal, Canada, between October 1985 and November 2006. Our cohort includes 189 boys and 132 girls with a median age of 4.7 years, all French-Canadian from the province of Quebec, Canada. Parental DNA was available for 203 of the probands. Healthy controls ($n=329$) consisted of French-Canadian individuals recruited while using clinical departments other than Hematology-Oncology of the Sainte-Justine Hospital.

SNPs, genotyping and quality control checks. We selected 29 SNPs from 12 candidate cell cycle genes for the analysis (Table 4). Genes were selected

based on their function in regulating the G1/S cell cycle checkpoint. Regulatory SNPs (found to lie within the proximal promoter region) were chosen based on the hypothesis that variation in gene dosage of such critical cell cycle genes due to functional regulatory polymorphisms could influence cancer susceptibility by altering cell homeostasis (28). For the purpose of this study using a French-Canadian cohort, we considered European-specific SNPs previously identified in (29). DNA was isolated from buccal epithelial cells, peripheral blood or bone marrow in remission as previously described (30). SNPs were genotyped using the Luminex xMAP/Autoplex Analyser CS1000 system (Perkin Elmer, Waltham, MA). Genetic variants were amplified using allele-specific primer extension in multiplexed assays and hybridized to Luminex MicroPlex TM –xTAG Microspheres as per Koo et al. (31). Primer sequences, amplification conditions, and reaction conditions are available upon request. Genotypes were called using the Automatic Luminex Genotyping (ALG) software (32). Three negative controls and three sample duplicates were used on each 96-well DNA plate. The average genotype call rate was 99.8% and rates of discordance were below 3.3%. In addition, Hardy-Weinberg equilibrium was tested using the χ^2 goodness of fit test and PedCheck (Version 1.1) was used to identify genotype incompatibilities using the familial data (33); inconsistent case-parent trios were removed from the analyses. Multiple testing corrections were performed on the single-step association tests using the Benjamini-Hochberg false discovery rate (FDR) method with a type I error rate of 10%; nominal P values are shown.

Results and Discussion

Although there are currently no data to document the frequency of events that lead to mating distortions in human populations, it is biologically plausible that MA might commonly occur. It is known that assortative mating (selection of a mate on the basis of phenotype leading to correlation between phenotypes of

mated individuals with respect to a given trait) can lead to genotype frequency differences between males and females (34-36). Other mechanisms leading to mating asymmetry however are unclear and their evolutionary consequences much less understood. If, for a specific marker, MA results in a departure from Hardy-Weinberg equilibrium, this genetic marker would be excluded from an association study following quality control. However MA could also arise through mating selection but with discrimination acting oppositely in each sex, or through differential individual mating success for the genotypes of each sex. Both these processes could lead to genotype frequency differences between sexes that would not lead to detectable deviations from Hardy-Weinberg equilibrium but that may incur important biases in fetomaternal association testing. In addition to these biological causes, low levels of MA could arise in a study sample simply due to the sampling process. Although parents of controls allow direct testing for bias due to MA in fetomaternal genotype association testing (9), these samples are difficult to collect and a method that can combine case-control and family-based data and provide a valid analytical framework for fetomaternal association testing in the presence of MA is currently not available.

Simulation study

In this study we used simulations to investigate the ability of three fetomaternal genotype association tests: 1) the log-linear, likelihood-based method of Weinberg and Umbach (9) using a case-parent/case-control design (HD-NPC), 2) the conditional logistic regression approach of Cordell et al. (7) combined with a case-control test (CC+CLR), and 3) Weinberg and Umbach's hybrid design using log-linear regression analysis, to distinguish between offspring and maternal genetic contributions to disease in the presence of MA. It should be noted that since both the log-linear and CLR frameworks are equivalent in terms of power and type I error for the detection of maternal genotype effects (7), our conclusions regarding the CC+CLR approach would also apply to a case-control

combined with the log-linear linear framework of Wilcox et al. (1) and Weinberg et al. (3) using case-parent triads. We evaluated type I error rates and power of the methods under varying degrees of MA, and genotypic risk models involving child, mother or joint effects of both child and mother (Table 2). For clarity reasons, we present the results for multiplicative genotype effect models only. However similar results were obtained under dominant and recessive models, with recessive models yielding expected decreases in power across all methods, particularly at low allele frequencies.

As expected, all three methods showed similar low type I error rates, around 5% and similar power, above 80% for the detection of child effects ($GC_{11} = 1$; $GC_{12} = 2$; $GC_{22} = 3$) under MS as well as under MA (data not shown). For a maternal main effect ($GM_{11} = 1$; $GM_{12} = 2$; $GM_{22} = 3$), type I error and power under MS were also within the acceptable ranges (data not shown). By contrast, under MA the CC+CLR method yielded unacceptably high type I error rates for the Mother test (Figure 2A). Although we expected that the method developed by Cordell et al. would be susceptible to the confounding incurred by MA, we found that CLR does not withstand even low levels of asymmetry ($\Delta C \sim 0.1$) so that even weak assumptions concerning population distributions of parental genotypes could lead to important bias.

The validity of the Mother tests for HD and HD-NPC were unaffected by MA, with type I error rates below the 5% threshold (Figure 2A). Power of the HD design was maintained at 100% and was unaffected by MA under the simulation conditions considered here, whereas power of HD-NPC considerably dropped, averaging around 30% (Figure 2B). When HD-NPC is used under asymmetry, genotypes for the parents of controls are inferred based on mating-type frequencies estimated from the parents of the cases and based on the assumptions that the control offspring genotypes follow Mendelian proportions

in relation to their parents (9). Hence, the maternal effect present in the case triads is partially captured in the inference of the mating-type frequencies for the parents of controls, resulting in a loss of power to detect this maternal effect as it becomes confounded with the estimated asymmetry.

The stepwise procedure allows maternal and case effects to be distinguished by estimating maternal effects independently of offspring effects and provides a valid test for joint fetomaternal associations. Under the null model and MS, the stepwise likelihood-ratio testing procedure yielded type I error rates close to 0% (data not shown), most likely due to the over-conservative Bonferroni correction that was applied. Since the Mother and Child tests are not completely independent a permutation test procedure would provide less-conservative type I error estimates. When we modelled multiplicative effects of both offspring and maternal effects ($GC_{11} = GM_{11} = 2$ and $GC_{12} = GM_{12} = 3$), the power to detect fetomaternal associations using the forward stepwise procedure was comparable for all three methods and increased with increasing allele prevalence for each method, reaching 80% for MAFs > 0.20 (Figure 3A).

We then evaluated type I error rates and power for the stepwise procedure in the presence of MA. The performance of both HD and HD-NPC was unaffected by MA with type I error rates close to zero even in the presence of high levels of asymmetry (Supplementary Figure 2). The type I error rate of the CC+CLR approach was close to 5% (Supplementary Figure 2) given that both child and maternal effects had to be falsely detected in order for the replicate to be counted as a false-positive and the case-control component of the test was robust against spurious child associations. Because the CLR maternal test is not valid under even low levels of MA, we assessed the power of the other two approaches to detect fetomaternal associations for varying levels of MA (Figure 3B). HD performed significantly better than HD-NPC: sensitivity of the HD

design averaged around 100% whereas power of HD-NPC was again significantly lower, averaging around 40%.

Based on the above findings, no method seemed to provide a net advantage under MS for these simulation conditions; nor did we observe any significant loss of power or robustness when the conditional logistic regression or the log-linear, likelihood-based approaches were used without parents of controls. Although power was significantly reduced, the log-linear, likelihood-based approach using controls rather than parents had little effect on the specificity of the association tests in the presence of MA. However, for the analyses performed on MA simulated datasets, we forced the estimation of nine mating-type parameters. In practice, no a priori assumptions regarding MS could be made. To verify the robustness of both methods for the detection of maternal (and fetomaternal) genotype effects in the presence of asymmetry, we measured type I error rates for the Mother test for scenarios in which either MA or MS models are assumed, and for a scenario in which no a priori hypothesis is made but rather MS is first evaluated in a three degree-of-freedom likelihood ratio test and the appropriate models (MA or MS) are subsequently used for association testing. These results show that if one assumes MS and this assumption is violated, type I error rates for the maternal test for both HD-NPC and HD are significantly inflated (Figure 4). However, first testing for asymmetry and then adjusting the association analyses accordingly provides accurate type I error rates for both methods. Similar results were obtained for the forward stepwise procedure (data not shown).

The specificity of the HD-NPC test therefore relies on its ability to detect MA and then use mating-type models accounting for asymmetry to test for association. Given that there are no biological references for the amount of MA that occurs in human populations, we evaluated the capacity of the HD and HD-NPC methods

to detect various levels of MA (Figure 5). Our simulation results showed that the power of HD-NPC to detect asymmetry above $\Delta C = 0.4$ was comparable to that of HD. For a risk allele frequency of $q = 0.3$, HD and HD-NPC reached 80% power at $\Delta C = 0.25$ and $\Delta C = 0.35$, respectively. Under low levels (ΔC ranging from 0 to 0.2) the sensitivity of both methods to detect asymmetry was threatened, especially for the HD-NPC approach (Figure 5). On the other hand, the lack of power of HD-NPC and HD to detect low levels of asymmetry is compensated by the fact that, without any a priori assumptions regarding mating symmetry, both methods maintained low type I error rates, at least under the simulation conditions presented here (Figure 4).

Therefore if asymmetry is not strong enough to be detected by the MA test it should not be falsely interpreted as a maternally-mediated effect. By contrast if a maternal effect is present and HD-NPC is used to test for mating asymmetry without parents of controls, type I error rates (of false detection of MA) are high (ranging from 0.8 to 0.9 for MAFs of 0.10 to 0.25, respectively) leading to a subsequent loss of power to detect maternal effects due to the over-parameterisation under the MA models (data not shown). Together these results show that the log-linear, likelihood-based stepwise procedure using unaffected offspring provides a valid framework to evaluate MS without leading to spurious maternal associations. And when parents of controls cannot be ascertained but an additional set of unrelated controls is available, one can safely use this approach to test for fetomaternal associations if willing to accept that certain confounded maternally-mediated effects may be missed when parental mating is asymmetric.

Fetomaternal association study of childhood acute lymphoblastic leukemia

Guided by our simulation results we went on to test for fetomaternal associations between 29 SNPs in the proximal promoter regions of 12 cell cycle genes (29) and the susceptibility to childhood pre-B ALL. SNP frequencies were in agreement with those previously reported in other populations of European descent and all distributions were in Hardy-Weinberg equilibrium. Our dataset consisted of 118 pre-B ALL patients, 203 ALL case-parent triads, and 329 unrelated controls. The lack of parents of controls prevents us from excluding MA in the source population. Based on the results from our simulation study, we used the log-linear framework to perform likelihood-based testing in a stepwise fashion. For each SNP, we performed a three degree-of-freedom likelihood-ratio test for asymmetry implemented in the LEM software using a slightly less stringent P value < 0.10 to reject symmetry in order to reduce false-positives in the tests for maternal effects. Under this threshold, we identified MA at variants rs1144944 ($P = 0.086$) and rs3730485 ($P = 0.095$) of the *MDM2* gene, as well as at *CDKN2B* variant rs2069416 ($P = 0.076$); we did not detect asymmetry at any of the remaining loci tested (P values > 0.10 ; data not shown). Consequently, MA models (nine mating-type parameters) were used to test for association at these three SNPs whereas MS models (six mating-type parameters) were used for the remaining 26 SNPs (see Figure 6, for the Child and Mother single-step test results and Supplementary Table 1 for complete likelihood-ratio chi-square test results). Nominally significant genotype associations at the level of the child were identified for *CDKN2A* rs36228834 (Child vs. Null; $P = 0.0007$), *CDKN1B* rs35756741 (Child vs. Null; $P = 0.0235$) and *CDKN2B* rs2069416 (Child vs. Null; $P = 0.0063$); however only *CDKN2A* rs36228834 and *CDKN2B* rs2069416 remained significant after multiple testing corrections (Supplementary Table 1). None of the other 26 loci revealed any significant child-mediated genetic associations with ALL and no significant maternal genotype effects were identified through the Mother vs. Null log-linear test (Figure 6 and Supplementary Table 1).

Nominally significant SNPs were further analyzed to detect putative joint fetomaternal effects. After accounting for the effect of the child's genotype, we found a significant maternal association at *CDKN2A* rs36228834 (Child+Mother vs. Child; $P = 0.0168$) only (Supplementary Table 1). The *CDKN2A* rs36228834 A allele was overrepresented in patients when compared with controls (genotype Fisher's exact $P = 0.005$) and carriers of the A risk allele were 2.5-fold more susceptible to ALL (Child vs. Null; TA vs. TT: OR= 2.48; 95%CI(1.45-4.15); TA/AA vs. TT: OR= 2.56; 95%CI(1.54-4.26)). This risk was further increased when the mother's genotype was included in the model (Child + Mother vs. Null; TA vs. TT: OR= 3.13; 95%CI(1.81-5.40); TA/AA vs. TT: OR= 3.20; 95%CI(1.85-5.53)) (Table 5). No further maternal association was detected for *CDKN1B* rs35756741 (Child + Mother vs. Child; $P = 0.59$) or *CDKN2B* rs2069416 (Child + Mother vs. Child; $P = 0.33$) (Supplementary Table 1). These results provide evidence of a novel fetomaternal effect at the *CDKN2A* rs36228834 locus that may influence pre-B ALL susceptibility among children, and a possible joint effect of both mother and child genotypes without main effects at *CDKN2B* rs36229158. Interestingly, although variant *CDKN2B* rs36229158 presented only a suggestive effect at the level of the child (Child vs. Null; $P = 0.06$), a significant association was found for the mother after we accounted for the genotype of the child (Child + Mother vs. Child; $P = 0.0217$) (Supplementary Table 1). Including the mother's genotype in the regression model significantly increased the risk 2.3-fold for carriers of a C allele (Child + Mother vs. Null; CT vs. CC: OR= 2.32; 95%CI(1.23-4.35); CT vs. CC: OR= 2.44; 95%CI(1.29-4.60)) (Table 5).

Independent replication is required in order to confirm the significance of these associations given that some of these variants did not withstand multiple testing correction (*CDKN1B* rs35756741 and *CDKN2B* rs36229158); and caution is warranted in the interpretation of the risk estimates as risk allele frequencies at loci *CDKN2A* rs36228834 and *CDKN2B* rs36229158 were low (MAFs ≤ 0.05),

yielding large confidence intervals, particularly for the rare homozygous genotype classes. We also recognize that the interpretation of our real-data results relies in part on our simulation results and there are certain limitations to our simulation study. These include the restricted number of models used in the method evaluations and the important assumptions of absence of population substructure and Hardy-Weinberg equilibrium, which necessitate further investigation. Our conclusions on the validity of the HD-NPC approach under MA should hold for other models. As we noted, the detection of asymmetry and maternal effects are partially confounded with HD-NPC and it is difficult to imagine a model of asymmetry for which differences in mating type frequencies would not be detected by the asymmetry test but would be captured by the Mother vs. Null test, thus leading to increased type I error. However, the reduction in power of HD-NPC compared to HD will likely be affected by the underlying genetic and asymmetry models and should be assessed under a wider range of models. Other important genetic effects should also be investigated, such as mother-gene child-gene interactions and parent-of-origin effects which are not addressed here but can also be involved in early-onset disorder risk.

Nonetheless, our results provide evidence that genes that regulate the cell cycle could play an important role during fetal development when the rate of cell growth and division is high both in child and mother. *In silico* analysis using the MatchTM software (37) revealed that all three variant loci lead to the disruption of putative transcription factor binding sites, including the loss of binding sites for FOS and MYB at *CDKN2A* rs36228834. The FOS oncoprotein stimulates transcription of genes containing AP-1 regulatory elements and may transform cells through alterations in DNA methylation and in histone deacetylation (38). Expression of FOS is 100-fold greater in human fetal membranes than in other normal human tissues and cells (39). The MYB transcription factor is essential for hematopoiesis and controls the proliferation and differentiation of

hematopoietic stem and progenitor cells (40). MYB is frequently involved in hematopoietic disorders including ALL (41). Although the biological relevance of our findings remains to be elucidated, our study suggests that promoter variation in the cell cycle inhibitor gene *CDKN2A*, and possibly *CDKN2B*, could disrupt transcription-factor binding and influence gene expression during gestation. Dysregulated cell division caused by aberrant cell cycle inhibitor gene expression in both mother and child could disrupt the maternal-fetal interface and affect important physiological processes such as the growth of the fetus and/or normal haematopoiesis and potentially lead to increased susceptibility to ALL.

Conclusions

Given the unique nature of childhood disorders, the investigation of parental genetics and maternally-contributed effects is a prerequisite not only for understanding disease etiology but also to pave the way toward new opportunities in preventive medicine. Although the most powerful approach is ideally the most desirable, in practice the best approach might be one that combines both valid detection of the possible underlying genetic associations involved in early-onset disorders and a feasible design in terms of ascertainment and genotyping costs. We have shown that the log-linear, likelihood-based framework using a case-triad/case-control design retains the ability to control for bias due to MA and can provide valid tests for maternally-contributed genotype effects even when the assumption of symmetry fails. Despite a modest sample size, we successfully used this approach to identify putative fetomaternal genotype effects in cell cycle inhibitor genes *CDKN2A* and *CDKN2B* that are associated with modified risks of childhood pre-B ALL. Although these genes have been previously associated with ALL (28), we have shown for the first time that their influence on ALL risk might be driven, in part,

by the maternal genotype. This study provides the first indication that maternal genotype effects can influence the risk of developing pediatric ALL, providing useful insights into the genetic mechanisms underlying this early-onset disease.

Figures

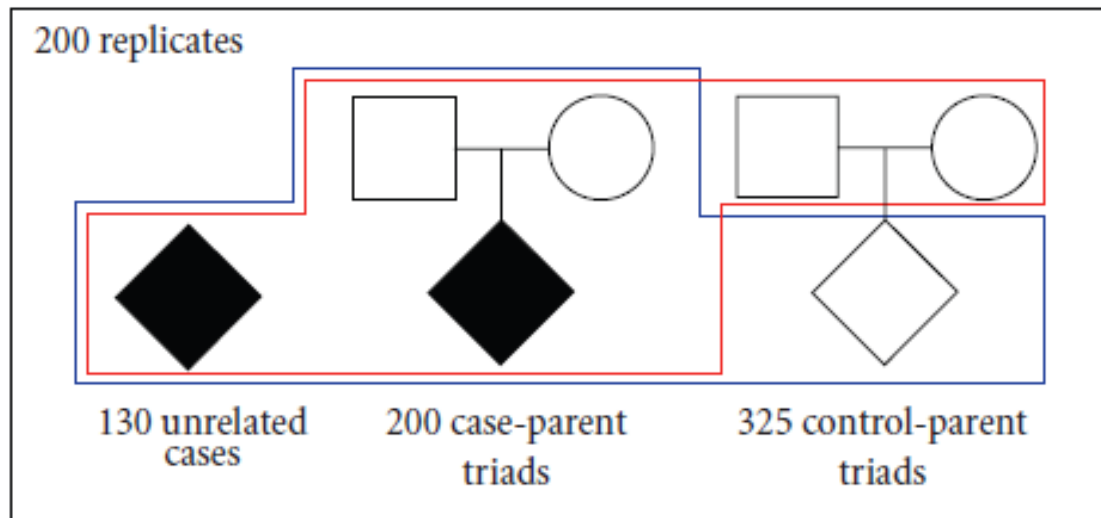


Figure 1. Cohort data used for simulations

200 replicate datasets were generated for each model. For all of our calculations we used the same overall number of affected individuals and case-parents however the methods differed in how the case and control data was utilized in each individual test (see Materials and Methods). The same number of affected individuals ($n = 330$) and case-parents ($n = 400$) were used under all three methods. For the combined case-control and conditional logistic regression analysis (CC+CLR) and the log-linear analysis using unrelated controls rather than their parents (HD-NPC), we used the genotypes of unrelated controls ($n = 325$, in blue), whereas for the hybrid design using parents of controls in a log-linear framework (HD), we incorporated the genotypic information of the parents of these controls rather than the controls themselves ($n = 650$, in red).

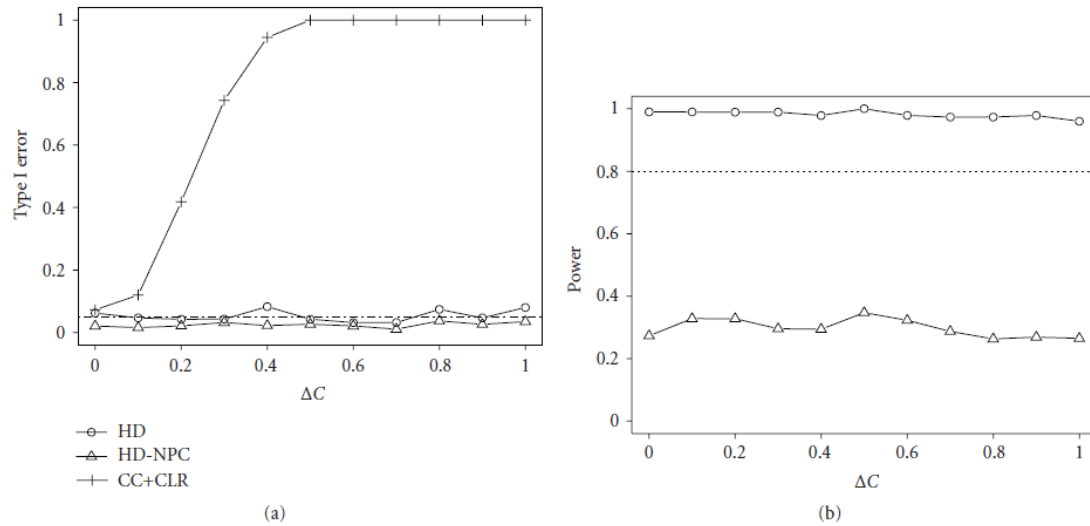


Figure 2. Type I error rates and power for the maternal association test under mating asymmetry

A) Type I error rates are given for the Mother vs. Null test as a function of departure from mating symmetry, as measured by ΔC , under the null model where $GC_{11} = GC_{12} = GC_{22} = GM_{11} = GM_{12} = GM_{22} = 1$ (Table 2, Model 5). Allele prevalence is set at $q = 0.3$. B) Power to detect a maternal effect is shown as a function of departure from mating symmetry, as measured by ΔC , for a scenario with multiplicative effects of the mother (Table 2, Model 7). Allele prevalence is set at $q = 0.3$. MA models (9 mating-type parameters) were used for log-linear regression under the assumption of mating asymmetry. Horizontal reference lines indicate type I error rate of $\alpha = 0.05$ (A) and power = 0.8 (B). HD, hybrid design using parents of controls in a log-linear framework; HD-NPC, log-linear analysis using unrelated controls rather than their parents; CC+CLR, combined case-control and conditional logistic regression analysis.

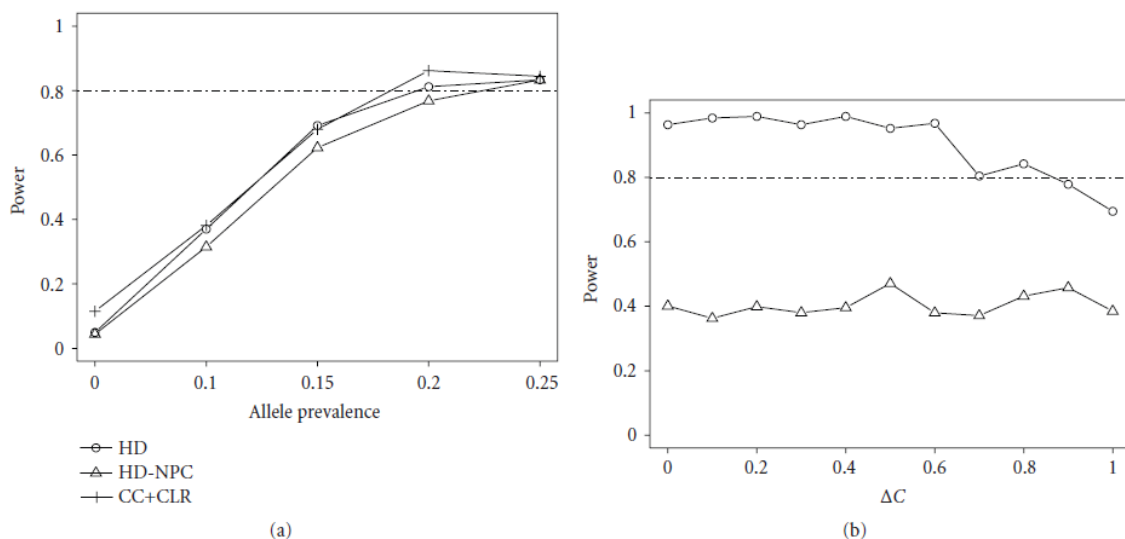


Figure 3. Power of the forward stepwise procedure to detect joint fetomaternal associations

A) Power for the HD, HD-NPC and CC+CLR methods is shown as a function of allele prevalence for a scenario with mating symmetry and multiplicative effects of both Child and Mother (Table 2, Model 4). B) Power for HD and HD-NPC is shown as a function of departure from mating symmetry, as measured by ΔC , for a scenario with mating asymmetry and multiplicative effects of both Child and Mother (Table 2, Model 8). Allele prevalence is set at $q = 0.3$. MA models (9 mating-type parameters) were used for log-linear regression under the assumption of mating asymmetry. Horizontal reference lines indicate power = 0.8. HD, hybrid design using parents of controls in a log-linear framework; HD-NPC, log-linear analysis using unrelated controls rather than their parents; CC+CLR, combined case-control and conditional logistic regression analysis.

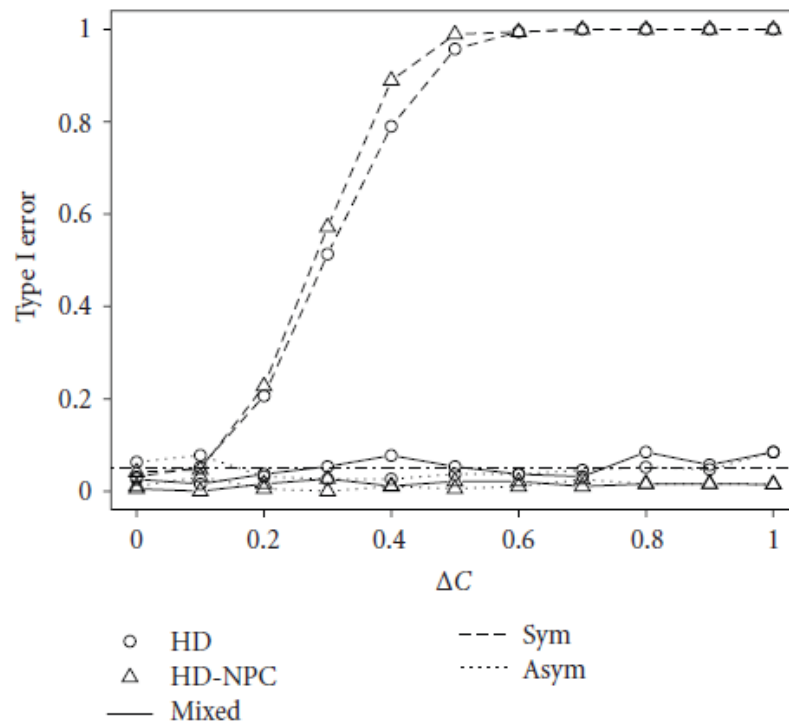


Figure 4. Type I error rate for the HD and HD-NPC approaches for the maternal association test assuming either mating symmetry or asymmetry

Type I error rates are shown as a function of departure from mating symmetry, as measured by ΔC , under the null model where $GC_{11} = GC_{12} = GC_{22} = GM_{11} = GM_{12} = GM_{22} = 1$ (Table 2, Model 5). Allele prevalence is set at $q = 0.3$. Log-linear, likelihood-based testing (Mother vs. Null) was performed assuming either mating asymmetry estimating nine mating-type parameters in the likelihood-ratio tests (Asym), or mating symmetry using only six mating-type parameters in the likelihood-ratio tests (Sym), or with no a priori hypothesis, using a 3 degree of freedom likelihood-ratio test to first test the assumption of mating symmetry (Mating Asymmetry (9df) vs. Mating Symmetry (6df)) and then using either the mating asymmetry models when asymmetry was detected or the mating symmetry models when it was not (Mixed). Horizontal reference line indicates type I error rate of $\alpha = 0.05$. HD, hybrid design using parents of controls in a log-linear framework; HD-NPC, log-linear analysis using unrelated controls rather than their parents.

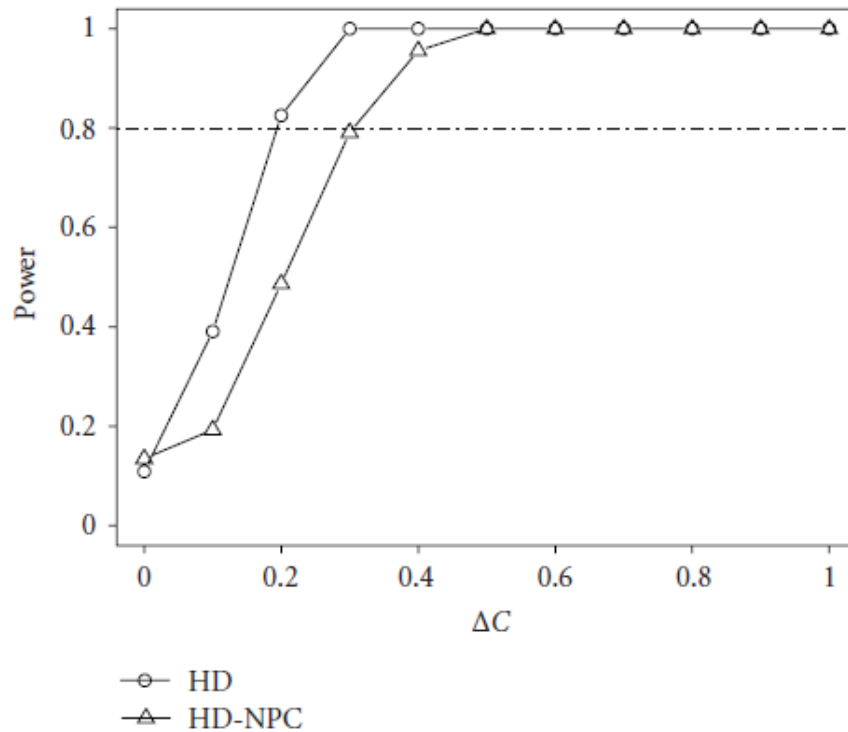


Figure 5. Power of HD and HD-NPC to detect mating asymmetry

Power is shown as a function of departure from mating symmetry, as measured by ΔC , under the null model where $GC_{11} = GC_{12} = GC_{22} = GM_{11} = GM_{12} = GM_{22} = 1$ (Table 2, Model 5). Allele prevalence is set at $q = 0.3$. Mating asymmetry was evaluated in the log-linear framework using a 3 degree of freedom likelihood-ratio test comparing the 9 mating-type parameter model under MA to a 6 mating-type parameter model under MS. Horizontal reference line indicates power = 0.8. HD, hybrid design using parents of controls in a log-linear framework; HD-NPC, log-linear analysis using unrelated controls rather than their parents.

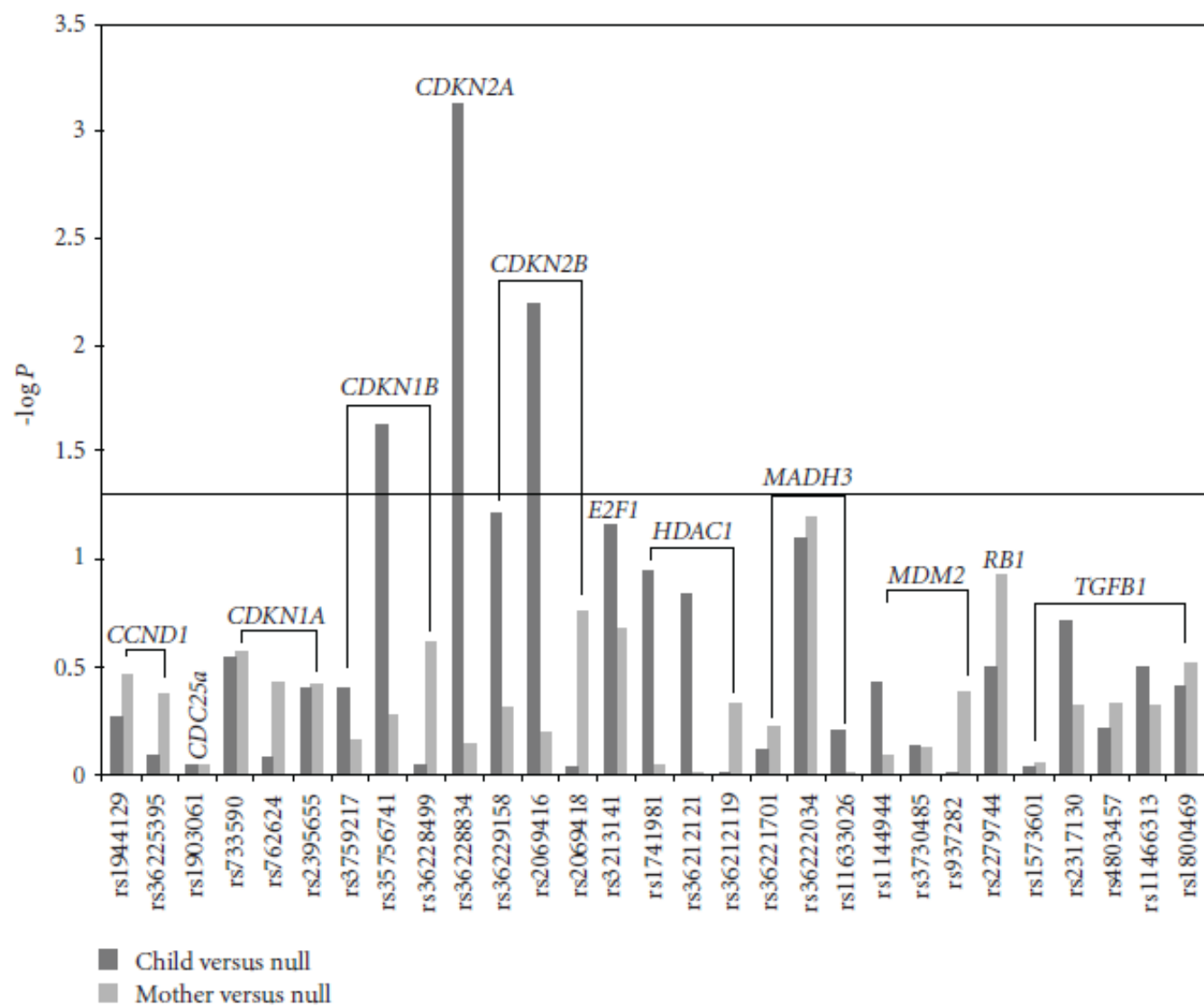
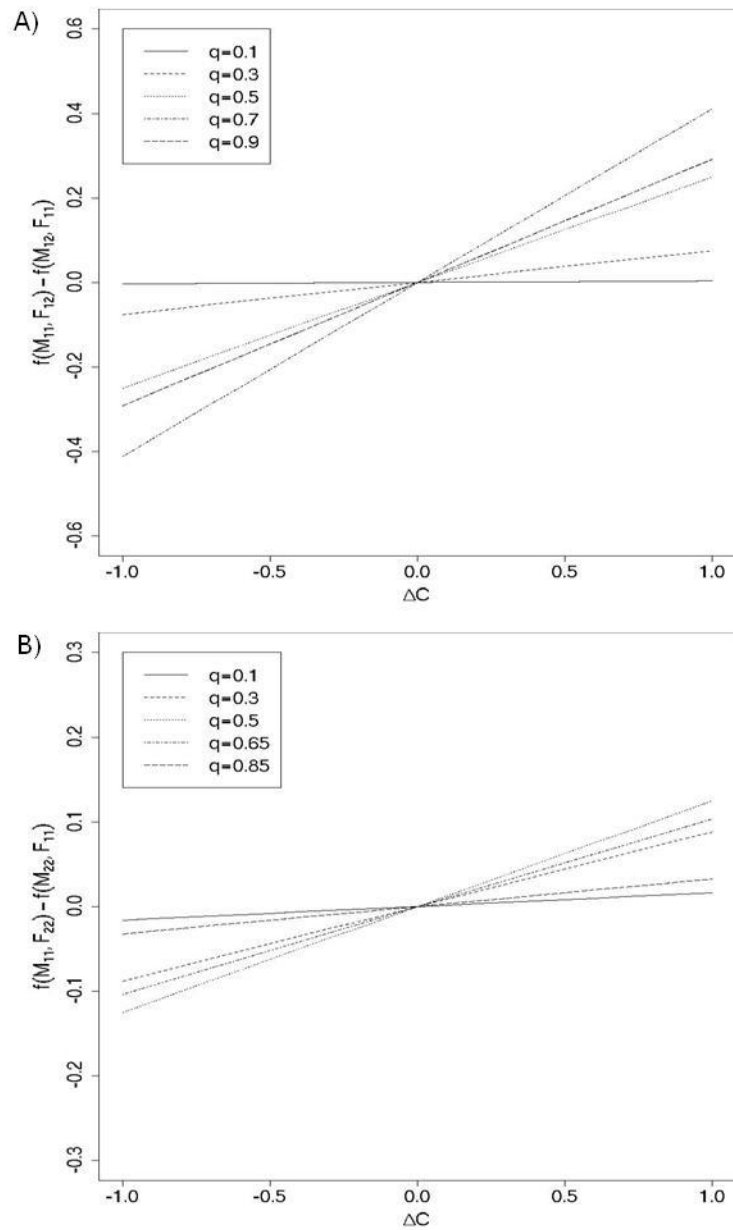


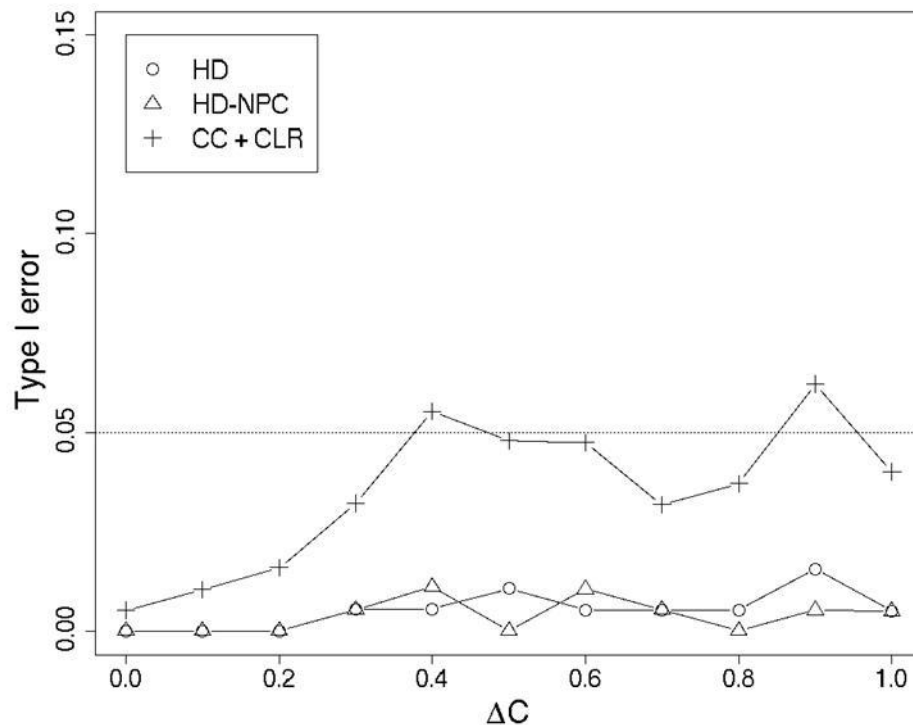
Figure 6. Log-linear, likelihood-ratio association analysis between 29 regulatory SNPs from 12 cell cycle genes and childhood pre-B acute lymphoblastic leukemia (ALL)

Log-linear regression analysis was performed in LEM using 203 case-triads, 118 unrelated ALL patients and 329 controls. Results of the likelihood-ratio chi-square tests ($-\log P$) are shown for the single-step Child (blue) and Mother (red) tests. Mating symmetry (i.e. six mating-type parameters) was assumed at all loci but variants rs1144944 and rs3730485 of the *MDM2* gene, as well as *CDKN2B* rs2069416 for which MA models (nine mating-type parameters) were used to test for association. Horizontal reference line indicates P value of 0.05. Note that for the tri-allelic variant *CDKN2B* rs2069416 (A>T,C) individuals were grouped according to their T allele such that $** > *T > TT$. See Supplementary Table 1 for complete likelihood-ratio chi-square test results.



Supplementary Figure 1. Relative distribution of reciprocal mating types as a function of mating asymmetry

Genotype frequency differences between mating type pairs are given as a function of ΔC , the measure of the departure from mating symmetry. Frequency differences are shown for Mother-Father mating type 1 (M_{11}, F_{12} vs. M_{12}, F_{11}) in A) and mating type 2 (M_{11}, F_{22} vs. M_{22}, F_{11}) in B) for allele frequencies (q) ranging from 0.1 to 0.9.



Supplementary Figure 2. Type I error rates for fetomaternal association testing under mating asymmetry

Type I error rates are given for the stepwise procedure and are shown as a function of departure from mating symmetry, as measured by ΔC , under the null model where $GC_{11} = GC_{12} = GC_{22} = GM_{11} = GM_{12} = GM_{22} = 1$ (Table 2, Model 5). Allele prevalence is set at $q = 0.3$. MA models (9 mating-type parameters) were used for log-linear regression under the assumption of mating asymmetry. Horizontal reference lines indicate type I error rate of $\alpha = 0.05$. HD, hybrid design using parents of controls in a log-linear framework; HD-NPC, log-linear analysis using unrelated controls rather than their parents; CC+CLR, combined case-control and conditional logistic regression analysis.

Tables

Table 1. Forward stepwise likelihood-ratio testing procedure used to dissect child and maternal genotype associations

Weinberg and Umbach log-linear approach using case-triads, unrelated cases and unrelated controls (HD-NPC) or parents of controls (HD)	Genotypic case-control test combined with the conditional logistic regression approach of Cordell et al. using case-triads (CC+CLR)
<i>Step 1.</i>	<i>Step 1.</i>
Child genotypic effect	Child genotypic association test (CC)
Null versus CG LR Chi-square (2df)	Chi-square or Fisher's exact test (2df)
And	And
Maternal genotypic effect	Maternal genotypic effect (CLR)
Null versus MG LR Chi-square (2df)	Null versus MG LR Chi-square (2df)
<i>Step 2. (depending on Step 1)</i>	<i>Step 2. (depending on Step 1)</i>
Maternal effect given child effect	Maternal effect given child effect (CLR)
CG versus CG+MG LR Chi-square (2df)	CG versus CG+MG LR Chi-square (2df)
Or	Or
Child effect given maternal effect	Child effect given maternal effect (CLR)
MG versus CG+MG LR Chi-square (2df)	MG versus CG+MG LR Chi-square (2df)

Likelihood-ratio tests were performed in a forward stepwise fashion. The most significant single-step test (Child vs. Null or Mother vs. Null) was tested against a joint effects model in a 2 degree of freedom likelihood-ratio test (Child + Mother vs. Child or Child + Mother vs. Mother). LR Chi2 indicates likelihood-ratio chi-square test; df, degrees of freedom; GC Child genotype relative risk; GM, Mother genotype relative risk.

Table 2. The eight simulation models used for evaluation of the fetomaternal association tests

Model	Child effect	Mother effect	MAF	ΔC
1	—	—	0.05 to 0.25	MS
2	$GC_{11} = 1; GC_{12} = 2; GC_{22} = 3$	—	0.05 to 0.25	MS
3	—	$GM_{11} = 1; GM_{12} = 2; GM_{22} = 3$	0.05 to 0.25	MS
4	$GC_{11} = 1; GC_{12} = 2; GC_{22} = 3$	$GM_{11} = 1; GM_{12} = 2; GM_{22} = 3$	0.05 to 0.25	MS
5	—	—	0.3	0 to 1
6	$GC_{11} = 1; GC_{12} = 2; GC_{22} = 3$	—	0.3	0 to 1
7	—	$GM_{11} = 1; GM_{12} = 2; GM_{22} = 3$	0.3	0 to 1
8	$GC_{11} = 1; GC_{12} = 2; GC_{22} = 3$	$GM_{11} = 1; GM_{12} = 2; GM_{22} = 3$	0.3	0 to 1

(—): indicates a null risk model where the genotype relative risks (GRRs) are $GRR_{11} = GRR_{12} = GRR_{22} = 1$. GC indicates Child genotype relative risk; GM, Mother genotype relative risk; MAF, Minor allele frequency; ΔC , mating-pair disequilibrium. MS indicates mating symmetry where $\Delta C = 0$.

Table 3. Parental genotype distributions under mating symmetry and mating asymmetry

Mating type	Parental genotypes (Mother-Father)	Expected distribution	
		Mating symmetry	Mating asymmetry
0	11-11	$N * \mu_0$	$N * \mu_0$
1	11-12	$N * \mu_1$	$(2 - C_1) * N * \mu_1$
	12-11	$N * \mu_1$	$C_1 * N * \mu_1$
2	11-22	$N * \mu_2$	$(2 - C_2) * N * \mu_2$
	22-11	$N * \mu_2$	$C_2 * N * \mu_2$
3	12-12	$N * \mu_3$	$N * \mu_3$
4	12-22	$N * \mu_4$	$(2 - C_4) * N * \mu_4$
	22-12	$N * \mu_4$	$C_4 * N * \mu_4$
5	22-22	$N * \mu_5$	$N * \mu_5$

Alleles for a biallelic locus are denoted 1 and 2 and the corresponding genotypes 11, 12 or 22. N indicates the number of individuals in the sample; μ_i the i^{th} mating type probability under the assumption of random mating; C_i the mating-pair disequilibrium for the i^{th} parental couple. C is a multiplicative factor between 0 and 2 that describes the over representation (>1), under representation (<1) or symmetry ($=1$) of a mate-pair combination in the corresponding i^{th} mating type.

Table 4. Genes and DNA variants genotyped in the pre-B ALL association study

Gene (Chromosome)	DNA variant	Position	MAF
<i>CCND1</i> (11q13)			
	rs1944129	69,163,116	0.4876
	rs36225395	69,163,517	0.4523
<i>CDC25a</i> (3p21)			
	rs1903061	48,206,923	0.1028
<i>CDKN1A</i> (6p21)			
	rs733590	36,753,181	0.3616
	rs762624	36,753,566	0.2714
	rs2395655	36,753,674	0.3968
<i>CDKN1B</i> (12p13)			
	rs3759217	12,759,719	0.1159
	rs35756741	12,759,968	0.0865
	rs36228499	12,761,203	0.4342
<i>CDKN2A</i> (9p21)			
	rs36228834	21,965,319	0.0512
<i>CDKN2B</i> (9p21)			
	rs36229158	22,000,681	0.0282
	rs2069416	22,000,004	0.3742/0.0271
	rs2069418	21,999,698	0.4272
<i>E2F1</i> (20q11)			
	rs3213141	31,738,041	0.2405
<i>HDAC1</i> (1p35)			
	rs1741981	32,529,026	0.3302
	rs36212121	32,529,102	0.0031
	rs36212119	32,529,840	0.0846
<i>MADH3</i> (15q22)			
	rs36221701	65,143,543	0.1199
	rs36222034	65,144,732	0.1111
	rs11633026	65,144,812	0.1235

<i>MDM2</i> (12q15)			
	rs1144944	67,486,752	0.4954
	rs3730485	67,487,073– 67,487,112	0.4052
	rs937282	67,488,064	0.483
	rs2279744	67,488,847	0.3662
<i>RB1</i> (13q14)			
	rs1573601	47,774,358	0.2484
<i>TGFB1</i> (19q13)			
	rs2317130	46,553,514	0.3141
	rs4803457	46,553,199	0.3937
	rs11466313	46,553,177– 46,553,178	0.3096
	rs1800469	46,552,136	0.3127

DNA variant positions relative to dbSNP build 130. MAF indicates minor allele frequency and was calculated on a control cohort consisting of 329 healthy individuals of European descent.

Table 5. Distribution of *CDKN2A* rs36228834 and *CDKN2B* rs36229158 genotypes and associated risk estimates for pre-B ALL susceptibility among children

Gene, DNA variant, and genotype		No. (%)			Log-linear regression analysis			
	ALL patients	ALL mothers	ALL fathers	Controls	Model	Genotype	Child OR (95% CI)	P
CDKN2A								
rs36228834					Child versus Null	TA versus TT	2.48 (1.45–4.15)	.001
TT	266 (86.6)	160 (93.0)	149 (86.6)	298 (93.7)		AA versus TT	9.87 (0.89–109.69)	
TA	39 (12.7)	12 (7.0)	22 (12.8)	19 (6.0)		TA/AA versus TT	2.56 (1.54–4.26)	<.0005
AA	2 (0.7)	0	1 (0.6)	1 (0.3)	Child + Mother versus Null	TA versus TT	3.13 (1.81–5.40)	<.0005
						AA versus TT	—	
						TA/AA versus TT	2.56 (1.54–4.26)	<.0005
CDKN2B								
rs36229158					Child versus Null	CT versus CC	1.77 (0.98–3.21)	.054
CC	277 (91.4)	164 (95.4)	155 (90.1)	302 (94.7)		TT versus CC	8.25 (0.75–91.3)	
CT	24 (7.9)	8 (4.6)	16 (9.3)	16 (5.0)		CT/TT versus CC	1.86 (1.04–3.34)	.037
TT	2 (0.7)	0	1 (0.6)	1 (0.3)	Child + Mother versus Null	CT versus CC	2.32 (1.23–4.35)	.033
						TT versus CC	—	
						CT/TT versus CC	2.44 (1.29–4.60)	.006

Percentages indicate number of individuals with a given genotype/total number of genotyped individuals. Risk estimation was performed using log-linear regression analysis as implemented in the LEM software. Child odd ratios were measured

using regression models consisting of the child genotype effect only (Child vs. Null) or both child and mother genotypes (Child + Mother vs. Null). Mating symmetry (i.e. six mating-type parameters) was assumed at both loci. P values of the Wald test provided by LEM are shown for either the 2 degree of freedom (2 child genotype effects) or 1 degree of freedom (1 child genotype effect resulting from the collapsed heterozygous/homozygous rare genotypes) tests. OR indicates odds ratio; CI, confidence interval.

Supplementary Table 1. Log-linear, likelihood-ratio association analysis between 29 regulatory SNPs from 12 cell cycle genes and childhood pre-B ALL

Gene, DNA variant	Model	LR Chi2 (df)	P
<i>CCND1</i>			
rs1944129	Child vs. Null	1.2295 (2)	0.5408
	Mother vs. Null	2.1512 (2)	0.3411
rs36225395	Child vs. Null	0.4281 (2)	0.8073
	Mother vs. Null	1.7239 (2)	0.4223
<i>CDC25a</i>			
rs1903061	Child vs. Null	0.2266 (2)	0.8929
	Mother vs. Null	0.2315 (2)	0.8907
<i>CDKN1A</i>			
rs733590	Child vs. Null	2.5003 (2)	0.2865
	Mother vs. Null	2.6508 (2)	0.2657
rs762624	Child vs. Null	0.3833 (2)	0.8256
	Mother vs. Null	1.9700 (2)	0.3734
rs2395655	Child vs. Null	1.8532 (2)	0.3959
	Mother vs. Null	1.9417 (2)	0.3788
<i>CDKN1B</i>			
rs3759217	Child vs. Null	1.8690 (2)	0.3928
	Mother vs. Null	0.7517 (2)	0.6867
rs35756741	Child vs. Null	7.5044 (2)	0.0235
	Mother vs. Null	1.2860 (2)	0.5257
	Child + Mother vs. Null	8.5604 (4)	0.0731
	Child + Mother vs. Child	1.0560 (2)	0.5898
rs36228499	Child vs. Null	0.2166 (2)	0.8973
	Mother vs. Null	2.8311 (2)	0.2428
<i>CDKN2A</i>			
rs36228834	Child vs. Null	14.4130 (2)	0.0007
	Mother vs. Null	0.65476 (2)	0.7209
	Child + Mother vs. Null	22.5809 (4)	0.0002
	Child + Mother vs. Child	8.1678 (2)	0.0168
<i>CDKN2B</i>			
rs36229158	Child vs. Null	5.6281 (2)	0.0600

	Mother vs. Null	1.4635 (2)	0.4811
	Child + Mother vs. Null	13.2885 (4)	0.0099
	Child + Mother vs. Child	7.6604 (2)	0.0217
rs2069416 ¹	Child vs. Null	10.1261 (2)	0.0063
	Mother vs. Null	0.9204 (2)	0.6312
	Child + Mother vs. Null	12.3208 (4)	0.0152
	Child + Mother vs. Child	2.1947 (2)	0.3338
rs2069418	Child vs. Null	0.1654 (2)	0.9206
	Mother vs. Null	3.4934 (2)	0.1744
E2F1			
rs3213141	Child vs. Null	5.3498 (2)	0.0689
	Mother vs. Null	3.1276 (2)	0.2093
HDAC1			
rs1741981	Child vs. Null	4.3556 (2)	0.1133
	Mother vs. Null	0.1855 (2)	0.9114
rs36212121	Child vs. Null	3.8760 (2)	0.1440
	Mother vs. Null	0.0251 (2)	0.9875
rs36212119	Child vs. Null	0.0436 (2)	0.9784
	Mother vs. Null	1.5479 (2)	0.4612
MADH3			
rs36221701	Child vs. Null	0.5317 (2)	0.7666
	Mother vs. Null	1.0401 (2)	0.5945
rs36222034	Child vs. Null	5.0971 (2)	0.0782
	Mother vs. Null	5.5164 (2)	0.0634
rs11633026	Child vs. Null	0.9712 (2)	0.6153
	Mother vs. Null	0.0025 (2)	0.9987
MDM2			
rs1144944	Child vs. Null	1.9874 (2)	0.3702
	Mother vs. Null	0.4247 (2)	0.8087
rs3730485	Child vs. Null	0.6397 (2)	0.7263
	Mother vs. Null	0.5579 (2)	0.7566
rs937282	Child vs. Null	0.0355 (2)	0.9824
	Mother vs. Null	1.7830 (2)	0.4100
rs2279744	Child vs. Null	2.3131 (2)	0.3146

	Mother vs. Null	4.2749 (2)	0.1180
RB1			
rs1573601	Child vs. Null	0.1689 (2)	0.9190
	Mother vs. Null	0.2512 (2)	0.8820
TGFB1			
rs2317130	Child vs. Null	3.3133 (2)	0.1908
	Mother vs. Null	1.4944 (2)	0.4737
rs4803457	Child vs. Null	1.0157 (2)	0.6018
	Mother vs. Null	1.5627 (2)	0.4578
rs11466313	Child vs. Null	2.3172 (2)	0.3139
	Mother vs. Null	1.4790 (2)	0.4773
rs1800469	Child vs. Null	1.8827 (2)	0.3901
	Mother vs. Null	2.4241 (2)	0.2976

Likelihood-ratio tests were performed in a forward stepwise fashion. The most significant single-step test (Child vs. Null or Mother vs. Null) was tested against a joint effects model in a 2 degree of freedom likelihood-ratio test (Child + Mother vs. Child or Child + Mother vs. Mother). Mating symmetry (i.e. six mating-type parameters) was assumed at all loci but variants rs1144944 and rs3730485 of the *MDM2* gene, as well as *CDKN2B* rs2069416 for which MA models (nine mating-type parameters) were used to test for association. Multiple testing corrections were performed on the single-step association tests using the Benjamini-Hochberg false discovery rate (FDR) method with a type I error rate of 10%; nominal P values are shown, those in bold remain significant after FDR adjustment. LR Chi2 indicates likelihood-ratio chi-square test; df, degrees of freedom.

¹*CDKN2B* SNP rs2069416 is tri-allelic A>T,C. For analysis, individuals were grouped according to their T allele such that **>*T>TT.

References

1. Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads". *Am J Epidemiol*. 1998 Nov 1;148(9):893-901.
2. Mitchell LE, Weinberg CR. Evaluation of offspring and maternal genetic effects on disease risk using a family-based approach: the "pent" design. *Am J Epidemiol*. 2005 Oct 1;162(7):676-85.
3. Weinberg CR, Wilcox AJ, Lie RT. A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet*. 1998 Apr;62(4):969-78.
4. Sinsheimer JS, Palmer CG, Woodward JA. Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test. *Genet Epidemiol*. 2003 Jan;24(1):1-13.
5. Weinberg CR. Studying parents and grandparents to assess genetic contributions to early-onset disease. *Am J Hum Genet*. 2003 Feb;72(2):438-47.
6. Palmer CG, Turunen JA, Sinsheimer JS, Minassian S, Paunio T, Lonnqvist J, et al. RHD maternal-fetal genotype incompatibility increases schizophrenia susceptibility. *Am J Hum Genet*. 2002 Dec;71(6):1312-9.
7. Cordell HJ, Barratt BJ, Clayton DG. Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epidemiol*. 2004 Apr;26(3):167-85.
8. Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet*. 2002 Jan;70(1):124-41.
9. Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am J Hum Genet*. 2005 Oct;77(4):627-36.
10. Shi M, Umbach DM, Vermeulen SH, Weinberg CR. Making the most of case-mother/control-mother studies. *Am J Epidemiol*. 2008 Sep 1;168(5):541-7.
11. Vermeulen SH, Shi M, Weinberg CR, Umbach DM. A hybrid design: case-parent triads supplemented by control-mother dyads. *Genet Epidemiol*. 2009 Feb;33(2):136-44.

12. Schaid DJ, Sommer SS. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *Am J Hum Genet.* 1993 Nov;53(5):1114-26.
13. Greaves M. Molecular genetics, natural history and the demise of childhood leukaemia. *Eur J Cancer.* 1999 Dec;35(14):1941-53.
14. Rowley JD. Molecular genetics in acute leukemia. *Leukemia.* 2000 Mar;14(3):513-7.
15. Wiemels JL, Cazzaniga G, Daniotti M, Eden OB, Addison GM, Masera G, et al. Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet.* 1999 Oct 30;354(9189):1499-503.
16. McHale CM, Wiemels JL, Zhang L, Ma X, Buffler PA, Guo W, et al. Prenatal origin of TEL-AML1-positive acute lymphoblastic leukemia in children born in California. *Genes Chromosomes Cancer.* 2003 May;37(1):36-43.
17. Greaves M. Infection, immune responses and the aetiology of childhood leukaemia. *Nat Rev Cancer.* 2006 Mar;6(3):193-203.
18. Greaves MF. Aetiology of acute leukaemia. *Lancet.* 1997 Feb 1;349(9048):344-9.
19. Chang JS, Selvin S, Metayer C, Crouse V, Golembesky A, Buffler PA. Parental smoking and the risk of childhood leukemia. *Am J Epidemiol.* 2006 Jun 15;163(12):1091-100.
20. Wen W, Shu XO, Potter JD, Severson RK, Buckley JD, Reaman GH, et al. Parental medication use and risk of childhood acute lymphoblastic leukemia. *Cancer.* 2002 Oct 15;95(8):1786-94.
21. Infante-Rivard C, Siemiatycki J, Lakhani R, Nadon L. Maternal exposure to occupational solvents and childhood leukemia. *Environ Health Perspect.* 2005 Jun;113(6):787-92.
22. Infante-Rivard C, Sinnett D. Preconceptional paternal exposure to pesticides and increased risk of childhood leukaemia. *Lancet.* 1999 Nov 20;354(9192):1819.
23. Kwan ML, Metayer C, Crouse V, Buffler PA. Maternal illness and drug/medication use during the period surrounding pregnancy and risk of childhood leukemia among offspring. *Am J Epidemiol.* 2007 Jan 1;165(1):27-35.
24. Pombo-de-Oliveira MS, Koifman S. Infant acute leukemia and maternal exposures during pregnancy. *Cancer Epidemiol Biomarkers Prev.* 2006 Dec;15(12):2336-41.
25. Buyske S. Maternal genotype effects can alias case genotype effects in case-control studies. *Eur J Hum Genet.* 2008 Jul;16(7):783-5.

26. van Den Oord EJ, Vermunt JK. Testing for linkage disequilibrium, maternal effects, and imprinting with (In)complete case-parent triads, by use of the computer program LEM. *Am J Hum Genet.* 2000 Jan;66(1):335-8.
27. Krajcinovic M, Labuda D, Richer C, Karimi S, Sinnett D. Susceptibility to childhood acute lymphoblastic leukemia: influence of CYP1A1, CYP2D6, GSTM1, and GSTT1 genetic polymorphisms. *Blood.* 1999 Mar 1;93(5):1496-501.
28. Healy J, Belanger H, Beaulieu P, Lariviere M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood.* 2007 Jan 15;109(2):683-92.
29. Sinnett D, Beaulieu P, Belanger H, Lefebvre JF, Langlois S, Theberge MC, et al. Detection and characterization of DNA variants in the promoter regions of hundreds of human disease candidate genes. *Genomics.* 2006 Jun;87(6):704-10.
30. Baccichet A, Qualman SK, Sinnett D. Allelic loss in childhood acute lymphoblastic leukemia. *Leuk Res.* 1997 Sep;21(9):817-23.
31. Koo SH, Ong TC, Chong KT, Lee CG, Chew FT, Lee EJ. Multiplexed genotyping of ABC transporter polymorphisms with the Bioplex suspension array. *Biol Proced Online.* 2007;9:27-42.
32. Bourgey M, Larivière M, Richer C, Sinnett D. ALG: Genotype Calling of Luminex Arrays. (submitted). 2009.
33. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet.* 1998 Jul;63(1):259-66.
34. Kondrashov AS, Shpak M. On the origin of species by means of assortative mating. *Proc Biol Sci.* 1998 Dec 7;265(1412):2273-8.
35. Kirkpatrick M, Ravigne V. Speciation by natural and sexual selection: models and experiments. *Am Nat.* 2002 Mar;159 Suppl 3:S22-35.
36. de Cara MA, Barton NH, Kirkpatrick M. A model for the evolution of assortative mating. *Am Nat.* 2008 May;171(5):580-96.
37. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 2003 Jul 1;31(13):3576-9.
38. Bakin AV, Curran T. Role of DNA 5-methylcytosine transferase in cell transformation by fos. *Science.* 1999 Jan 15;283(5400):387-90.

39. Muller R, Verma IM, Adamson ED. Expression of c-onc genes: c-fos transcripts accumulate to high levels during development of mouse placenta, yolk sac and amnion. *EMBO J.* 1983;2(5):679-84.
40. Sandberg ML, Sutton SE, Pletcher MT, Wiltshire T, Tarantino LM, Hogenesch JB, et al. c-Myb and p300 regulate hematopoietic stem cell proliferation and differentiation. *Dev Cell.* 2005 Feb;8(2):153-66.
41. Clappier E, Cuccuini W, Kalota A, Crinquette A, Cayuela JM, Dik WA, et al. The C-MYB locus is involved in chromosomal translocation and genomic duplications in human T-cell acute leukemia (T-ALL), the translocation defining a new T-ALL subtype in very young children. *Blood.* 2007 Aug 15;110(4):1251-61.

Reprint Permissions



May 20, 2010

To whom it may concern:

This is to let you know that the Research Article titled "Detection of feto-maternal genotype associations in early-onset disorders: evaluation of different methods and their application to childhood leukemia," by Jasmine Healy, Mathieu Bourgey, Chantal Richer, D. Sinnett and Marie-Helene Roy-Gagnon has been accepted for publication on March 15, 2010, in Journal of Biomedicine and Biotechnology. The article will be published shortly. We hereby grant permission to use this article for no-profit use in Ms. Healy's dissertation.

Best regards,

Yours sincerely,

Miada

Miada Elsharkawy
Editorial Staff
Hindawi Publishing Corporation
<http://www.hindawi.com>

CHAPTER FIVE

*CONCLUSIONS
AND FUTURE PERSPECTIVES*

SUMMARY AND DISCUSSION OF MAIN FINDINGS

Unravelling the genetic architecture of complex diseases has proven to be a daunting task and current analytical approaches geared toward investigating high impact gene variations have provided few insights into the genetic complexities of polygenic diseases. The relative lack of success of these initial efforts emphasizes the need for study designs and methods that account for genetic effects other than those mediated by single, inherited genes. In this study I attempted to dissect the etiologic intricacies underlying childhood acute lymphoblastic leukemia (ALL) and identify genetic factors that shape interindividual variability in the susceptibility to this disease. ALL is the most common cancer among children and is likely caused by multiple genetic and environmental factors. However the identification of established genetic risk factors for ALL has been impeded by the complex and heterogenic nature of the disease.

To tackle this issue I used three approaches in my doctoral research project, as detailed in Chapters Two, Three, and Four. Using a unique study design consisting of both case-control and family-trio data, I performed a pathway-based candidate gene association study of childhood pre-B ALL. I selected genes based on their involvement in two core signalling pathways, the G1/S cell cycle checkpoint and DNA double-strand break repair (DSBR) pathways, following the hypothesis that deregulation of these mechanisms could lead to increased genomic instability and mutational burden and disrupt cell homeostasis, thereby influencing disease risk. To better depict the genetic foundation of childhood ALL I not only investigated the putative effects of individual genes on disease risk (Chapter Two) but attempted to further describe the combined effects of multiple genes through epistatic interactions or through whole pathway effects (Chapter Three). And for the first time, I assessed the

role of the mother's genes in modulating disease risk in her offspring, either directly or through joint effects with the child's genotype (Chapter Four). Through these studies I successfully identified genetic variants in both the cell cycle and DNA repair pathways that are associated with childhood ALL and provided evidence for the implication of these biologic mechanisms in leukemogenesis. Moreover, I identified genetic variants that act through the mother, at least in part, to influence the risk of ALL for the growing fetus; this is the first study to demonstrate the role of parental genetic contributions in pediatric cancers.

Novel susceptibility loci for childhood ALL that were identified through this study include cell cycle control genes *CDKN2A*, *CDKN2B*, and *HDAC1* and DNA double-strand break repair genes *BRCA1*, *XRCC4* and *XRCC5*. Individual SNPs and/or haplotypes in the promoter regions of these genes were shown to significantly modulate ALL risk. These data add to the growing list of genetic association studies that have identified susceptibility loci for childhood ALL (see Table 4, Chapter One). Yet for the most part the observed associations, including some of the ones identified through my studies, have been weak or inconsistent or have only been identified in a single study. Additional, larger studies are required to confirm the presumed genetic risk factors for ALL detected in my doctoral research and to validate their role in disease.

To this effect, we recently participated in a genome-wide association follow-up study in which 34 SNPs were replicated in additional case-control series (including the QcALL cohort) totaling 2,386 childhood ALL cases and 2,419 controls¹. Among the 34 SNPs interrogated, a single variant rs3731217 mapping to chromosomal region 9p21.3, was found to be strongly associated with

¹ (This work has recently been published (Variation in *CDKN2A* at 9p21.3 influences childhood acute lymphoblastic leukemia risk; Sherborne et al. *Nat Genet.* 2010;42(6):492-4; see Appendix IV for full manuscript).

childhood ALL within four replication series (1). The 9p21 region encompasses the *CDKN2A* (p16) and *CDKN2B* (p15) cell cycle inhibitor genes, the *CDKN2A* alternative reading frame transcript which encodes p14/ARF involved in activation of the tumor suppressor protein p53, the *MTAP* gene which encodes a phosphorylase that plays an important role in adenine and methionine metabolism, as well as the non-coding RNA *ANRIL*. This region is an important susceptibility locus for several diseases with a complex genetic background; variants within the region have been associated with coronary artery disease (2-4), ischaemic stroke (5), aortic aneurysm (6), type II diabetes (7, 8), glioma (9, 10), and malignant melanoma (11) through large-scale genome-wide investigation and candidate gene studies have reported SNPs in this region associated with breast (12, 13), ovarian (14), and pancreatic carcinoma (15), as well as melanoma (16).

While this GWA study (1) provides unequivocal evidence for a relationship between *CDKN2A* and childhood ALL risk, I am unable to correlate the GWAS findings with my candidate gene observations using my samples given that the genotype data from Sherborne et al. were not made available to us and the *CDKN2A* variant rs36228834 identified in my study has not been genotyped in publicly available datasets such as the HapMap or 1000 Genomes data. My data did indicate however that LD between variants *CDKN2A* rs36228834 and *CDKN2B* rs2069416 in my study is low ($D' = 0.81$, $R^2 = 0.02$) suggesting that they represent independent association signals. And based on data from European individuals genotyped in the 1000 Genomes Project, pairwise LD between variants *CDKN2A* rs3731217 and *CDKN2B* rs2069416 is also expected to be low ($D' = 0.32$ and $R^2 = 0.01$), further suggesting independence of the *CDKN2A* and *CDKN2B* associations.

The two *CDKN2A* SNPs lie 9.3kb apart on chromosome 9 and are not found within a region of particularly high LD (Figure 1). The variant reported by the GWAS (rs3731217) lies within intron 1 of the *CDKN2A/p14/ARF* alternative reading frame transcript, 9.3kb downstream from our association signal (rs36228834) which, although also in intron 1 of *CDKN2A/p14/ARF*, falls within the proximal promoter region of *CDKN2A/p16* (Figure 1). Both *CDKN2A* transcripts encode structurally and functionally different proteins. While Sherborne et al. found no SNPs in LD with variant rs3731217 in any of the coding regions of *CDKN2A* (including both p16 or p14 isoforms) or any of the other genes in the region, nor did they identify any regulatory regions surrounding rs3731217 or any effects on gene expression (1), *in vitro* data from my study showed differential DNA-protein binding within the promoter region of *CDKN2A/p16* and allele-specific promoter activity associated with variant rs36228834 (17). It is possible that these associations are independent, and it is also possible that they may even be linked to different genes within the 9p21.3 region (*CDKN2A/p16*, *CDKN2A/p14/ARF*, *CDKN2B*, *MTAP*, *ANRIL*). Given that we are currently unable to describe LD levels between these two loci, fine-mapping and additional functional studies are imperative in order to better define the causal basis of the 9p21.3 associations with childhood ALL.

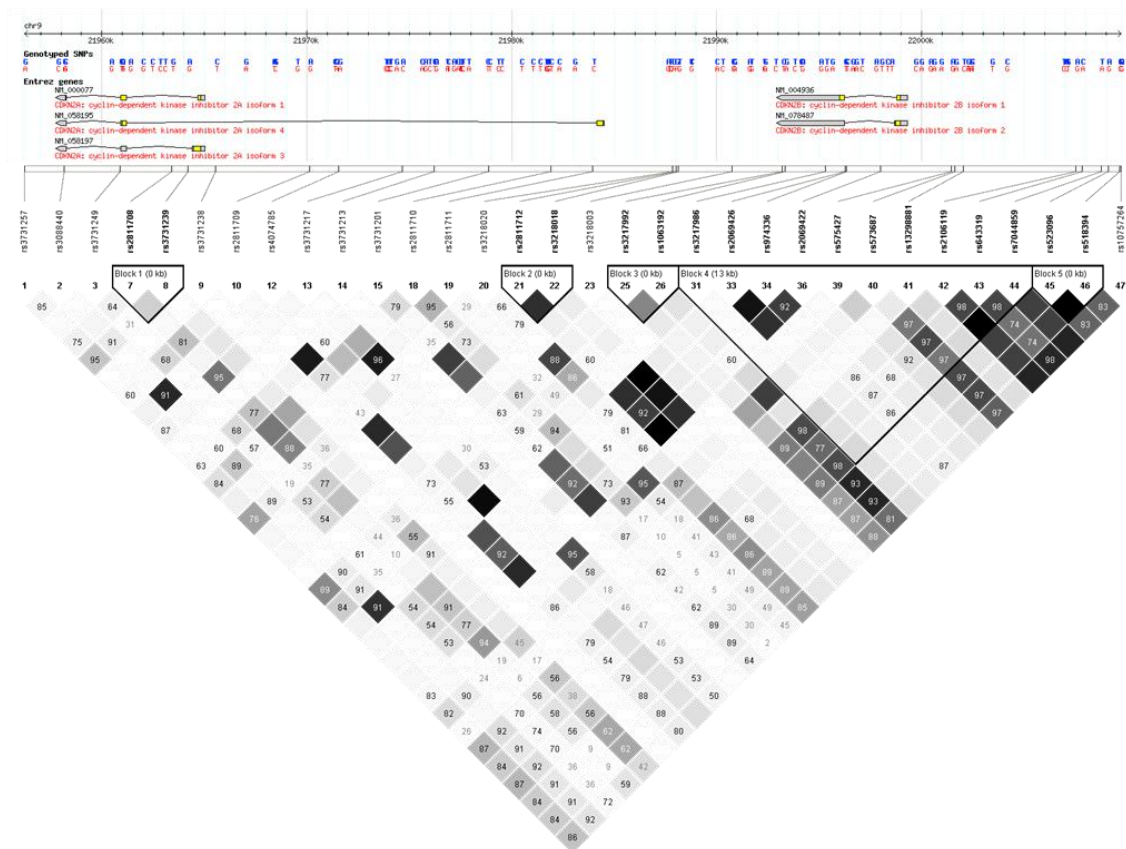


Figure 1. Details of the linkage disequilibrium region at 9p21.3

LD statistics (D' are shown, color scheme corresponds to r^2) from HapMap phase II CEU data are shown for a 60kb region (from 21,950,000 to 22,010,000 on NCBI build 36, dbSNP build 126) on chromosome 9p21.3 containing the *CDKN2A* (isoforms 1 (p16), 3, and 4 (p14)) and *CDKN2B* (isoforms 1 and 2) genes, as well as the non-coding RNA *ANRIL* (not shown). The darker shading indicates strong LD between SNPs. The *CDKN2A* association signal reported in Sherborne et al. (1) (rs3731217) corresponds to SNP 13. Given that variants *CDKN2A* rs36228499 and *CDKN2B* rs2069416 identified in my study (17) have not been genotyped in the HapMap individuals, they are not shown here however they would lie between SNPs 8-9 (position 21,964,819) and 36-39 (position 21,999,754), respectively.

Although many forms of human cancers exist, each with their own complexities and distinct patterns of genetic aberrations, a restricted number of alterations are thought to be shared by most, if not all tumour types (18). Such mutations include those that disrupt cell cycle control checkpoints allowing uncontrolled proliferation and those that hinder DNA damage response leading to increased genomic instability. p53 plays a central role in orchestrating cell cycle arrest and/or apoptosis in response to DNA damage and oncogenic signalling (19). Though *p53* is mutated in over 50% of human cancers, somatic inactivation of this gene is infrequent in hematologic malignancies (20) suggesting that other mechanisms are likely responsible for disruption of these processes in leukemia. One such mechanism is through the inactivation of cell cycle inhibitor genes *CDKN2A* and *CDKN2B* which occurs frequently in childhood ALL either through large chromosomal deletions, promoter methylation or, less frequently, through point mutations (21-23). Our study was the first however to provide evidence of inherited germline variants in *CDKN2A* and *CDKN2B* associated with childhood ALL susceptibility (17). These cyclin-dependent kinase inhibitors (CDKIs) integrate mitogenic and growth inhibitory signals and coordinate passage through the G1/S cell cycle checkpoint. Our studies suggest that regulatory variants within the promoter regions of these genes could lead to disrupted transcription factor-binding and allele-specific differences in gene expression. Though the functional significance of the association results requires further *in vivo* validation, decreases in expression of *CDKN2A* and *CDKN2B* in hematopoietic cells due to the pSNPs identified in our study could lead to decreased cyclin D1-Cdk4/Cdk6 inhibition and subsequent hyperphosphorylation of the RB1 protein facilitating entry into S phase. Decreased CDKI levels could thereby provide a proliferative advantage to the hematopoietic cells and contribute to leukemogenesis. Reduced *CDKN2A* and *CDKN2B* expression levels could also contribute to increased genomic instability observed in leukemia by limiting the G1/S phase and the time allocated for proper DNA repair.

However, a recent study showed that the *cis*-acting transcriptional effects of disease-associated SNPs in the 9p21 region, including *CDKN2A* promoter variant rs36228834, strongly influenced the expression of the non-coding RNA *ANRIL* in an allele-specific manner, while associations with *CDKN2A* and *CDKN2B* expression were weaker and less consistent (24). It may very well be that the true 9p21 disease culprit is in fact *ANRIL*, the effects of which may be mediated by post-transcriptional regulation of *CDKN2A/2B*. From an evolutionary point of view, sequence variants acting in *cis* through *ANRIL* may reflect selection pressure against variants that have direct effects on the expression of critical gatekeeper genes such as *CDKN2A* and *CDKN2B*. Though future studies are warranted to fully describe the involvement of this region in the pathogenesis of childhood ALL, my results further corroborate the importance of genetic variation at 9p21 in complex disease susceptibility.

Among the other loci identified in my study, only *XRCC4* has previously been linked to childhood ALL, in a Taiwanese population (25). Reduced expression of double-strand break repair genes *XRCC4* and *XRCC5*, variants of which were also significantly associated with increased ALL risk in children in my study, caused by functional promoter SNPs could lead to decreased capacity of the nonhomologous end joining (NHEJ) DNA repair process in which *XRCC4* and *XRCC5* are involved (26) and lead to increased genomic instability and potentially leukemia. Sequence variation in the *XRCC4* and *XRCC5* genes have been linked to various other types of cancers (27-35) suggesting that reduced double-strand break repair caused by inherited polymorphisms in these genes could serve as a common mechanism for carcinogenesis.

Interestingly, double-strand break repair proteins *XRCC4* and *XRCC5* also play a central role in V(D)J recombination in B lymphocytes (36). Developing lymphocytes undergo somatic rearrangements through a series of DNA

recombination events (V(D)J recombination) to assemble a diverse repertoire of immunoglobins and antigen receptors (37, 38). This process requires introduction of double-strand breaks and subsequent repair through NHEJ (Figure 2). It is now clear that V(D)J recombination plays a role far beyond the generation of antigen receptor diversity and is also essential for precursor lymphocyte maturation and differentiation (38). In fact *XRCC4*-deficient mice experience defective lymphogenesis and embryonic lethality due to extensive apoptotic cell death; and while additional knockout of *p53* allows postnatal survival of *XRCC4*-deficient mice, these mice routinely succumb to pre-B cell lymphomas caused by impaired V(D)J recombination (39). Therefore, decreased expression levels of NHEJ repair genes *XRCC4* and *XRCC5* could lead to aberrant V(D)J recombination events in B-cell progenitors and produce oncogenic chromosomal translocations and the appearance of lymphocytic tumours. The t(1;19)/*E2A-PBX1* translocation, occurring in ~5% of pediatric B-cell ALLs, does have some characteristics of translocations mediated by V(D)J recombination including site-specific clustering of the breakpoints and the presence of additional nucleotides at the fusion junctions (40) (see Figure 2), yet the more frequent chromosomal translocations observed in B-cell precursor childhood ALL rarely involve V(D)J recombination segments (41). Sequence variation in the *XRCC4/5* genes could however contribute to aberrant V(D)J recombination in B cells and lead to accumulated DSBs and overall genomic instability and thus influence ALL susceptibility among children.

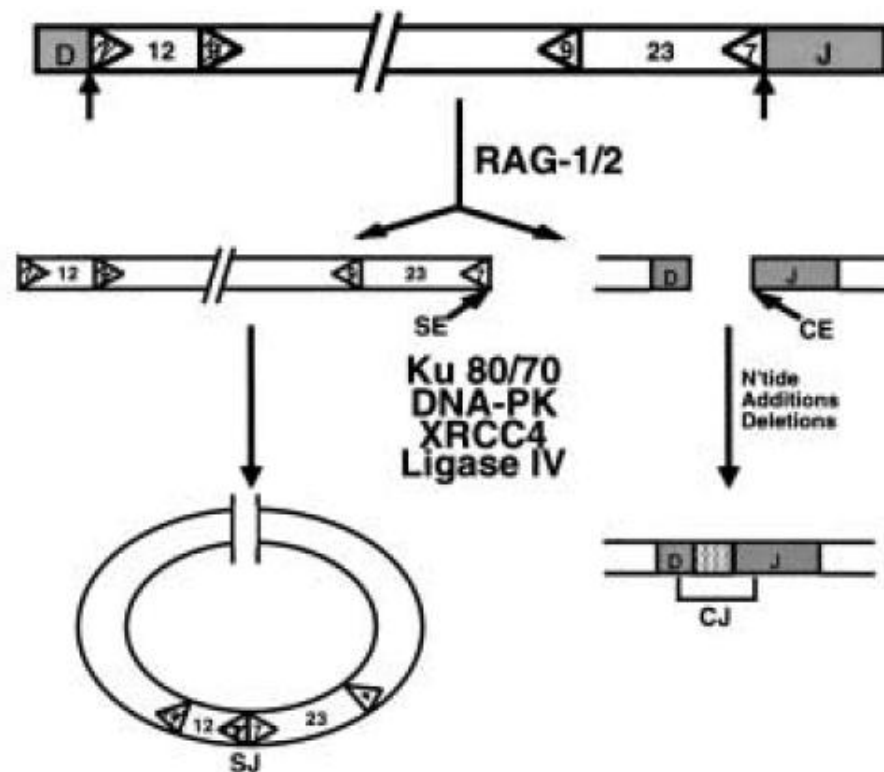


Figure 2. The V(D)J recombination mechanism

During development, B and T cells randomly select three types of gene segments, one V, one D and one J, which are represented by large arrays in mammalian genomes, and randomly fuse them by lymphoid-specific recombination events to produce variable coding exons. Shown here is a the fusion of D and J elements. The coding regions of the D and J gene segments are shown as shaded boxes and are flanked directly by recombinase recognition sites (RS; triangles). Initially the RAG complex introduces double-strand DNA breaks specifically at the RS/coding borders (arrows) of both participating gene segments to generate coding ends (CE) on the chromosome and signal ends (SE) as extrachromosomal DNA. Ubiquitous DNA repair proteins including XRCC5/6 (also known as Ku80/70) and XRCC4, ligate the SEs to produce a circular signal join (SJ). These SJs are deleted from the genome. CEs are further modified by nucleotide additions and/or deletions before ligation to increase the functional diversity of the resulting gene product known as the coding join (CJ).

Figure adapted from Oltz, 2001 (38).

This was the first study to identify germline variants in *BRCA1* associated with childhood ALL. As opposed to the more error-prone branch of the DSB repair pathway in which XRCC4 and XRCC5 are involved, *BRCA1* is involved in homologous recombination repair of DNA double-strand breaks (42). There is strong evidence for the implication of *BRCA1* deficiency in hematologic cancers: *BRCA1* is downregulated due to promoter hypermethylation in acute and chronic forms of myeloid leukemia (43, 44) and deleterious mutations in genes encoding proteins within the *BRCA1* repair pathway have been shown to increase risk for certain leukemias and lymphomas (44). While the mechanisms through which promoter sequence variation in *BRCA1* might influence ALL risk among children remain to be elucidated, it is possible that downregulation of *BRCA1* expression could lower fidelity of this error-free repair process and favour gene rearrangements characteristic of childhood ALL.

Finally, *HDAC1* was also identified as a potential susceptibility locus for childhood ALL. Histone acetylation/deacetylation is a major factor regulating chromatin structure and transcription and also plays an important role in controlling cell cycle progression (45). *HDAC1* physically interacts with RB1 and suppresses transcription of genes required for passage from G1 to S phase through histone deacetylation resulting in chromatin condensation and transcriptional repression (46). Tight control of *HDAC1* expression is therefore essential for normal progression through the cell cycle and altered activity could confer a growth advantage on hematopoietic cells by allowing expression of S-phase related genes or conversely by repressing expression of negative regulators of the cell cycle, and therefore contribute to leukemogenesis. Studies also suggest that *HDAC1* is involved in regulating hematopoietic progenitor cell differentiation; in fact *HDAC1* is overexpressed in acute myeloid leukemia cells where it contributes to leukemogenesis by halting myeloid differentiation (47). While it remains to be verified experimentally, we could speculate that regulatory polymorphisms in the *HDAC1* promoter could alter gene expression

levels and contribute to childhood ALL by disrupting normal lymphoid differentiation and by facilitating clonal expansion of the undifferentiated leukemic blasts. Moreover, widespread changes in gene expression are a feature of cancer and subtypes of ALL have been shown to have distinctive patterns of global gene expression in microarray analyses (48, 49). Epigenetic mechanisms, such as chromatin structural changes, promoter methylation and non-coding RNAs, play an important role in deregulating gene expression in cancer including leukemogenesis (50). Given the role of HDAC1 in histone deacetylation and chromatin remodelling, its altered activity could therefore contribute to global epigenetic regulation of gene expression in malignant hematopoiesis.

A Candidate Pathway Strategy

Dysregulation of biological processes that determine a cancer cell's fate is unlikely to result from the activity of a single gene but rather from multiple genes acting jointly in a common regulatory pathway or in parallel in concomitant pathways. The integrative pathway-based approach that I adopted in my study allowed biological and functional information pertinent to the genes and corresponding variants, to be directly incorporated into the association analysis (Chapter Three). While prior knowledge on pathway ontologies and *in silico* predictions of function and evolutionary conservation have previously been used in this manner (51, 52), this was the first time that a genetic association analysis attempted to take advantage of available data from functional assays to directly inform risk estimation. In the wake of what is now being called the “post-genomics era”, more and more genomic, proteomic, expression and other high-throughput data sources are becoming available. To make efficient use of this wealth of information in our quest to understanding complex disease processes, more comprehensive approaches such as the one used here, represent an interesting alternative to conventional methods of analysis. The use of

hierarchical models offers a unified approach to modelling the effects of entire pathways postulated to be relevant to disease, rather than look at the marginal effects of individual polymorphisms, and can act as a bridge between genetic epidemiology studies and laboratory data allowing multiple sources of prior information to be readily incorporated into the analytical framework. Traditional approaches considering one, or even two genes, at a time, neglect potential confounding effects and ignore any prior biological and functional knowledge that may be available. The hierarchical modeling approach also has the advantage of improving risk estimation and provides a means to dealing with the issue of multiple comparisons, reducing the likelihood of false positives. However given that only a limited number of genes within both pathways were interrogated in my study, further characterization of these gene networks, and other interconnected pathways such as xenobiotic metabolism and apoptosis, in childhood ALL development is required. And even though the biological model used in this study obviously represented an incomplete picture of the underlying disease process, it still provided insight into the mechanisms that may be driving leukemogenesis, suggesting a role for both the cell cycle and DNA double-strand break repair pathways, and allowed for individual gene variants associated with disease to be identified.

While their precise functional role in childhood ALL remains to be verified, my study also provided further evidence for the role of *cis*-acting regulatory variation in influencing disease susceptibility, potentially through modulated gene expression. Thus, my results suggest that germline variation in genes involved in the G1/S checkpoint and in DSBR may contribute to the increased genomic instability, the impaired differentiation, and the growth and/or survival advantage of hematopoietic cells which leads to leukemia. The novel susceptibility loci identified through my doctoral work provide additional insight into the biologic mechanisms underlying childhood ALL and perhaps other cancers as well (Figure 3). Replication of these genetic associations in independent childhood

ALL cohorts through our collaboration with the International Acute Lymphoblastic Leukemia Genetics Consortium, will help validate their role as childhood ALL susceptibility loci. This is especially important given that this was the first report of an association between these variants and childhood ALL and given the limited size and modest power of this study (see Appendix II). Furthermore, some of these associations, particularly *HDAC1*, were only weakly associated with disease. Further association testing in larger datasets and additional *in vitro* and *in vivo* functional assays will help identify the true causal variants at the identified loci and validate their implication in ALL and its particular subgroups.

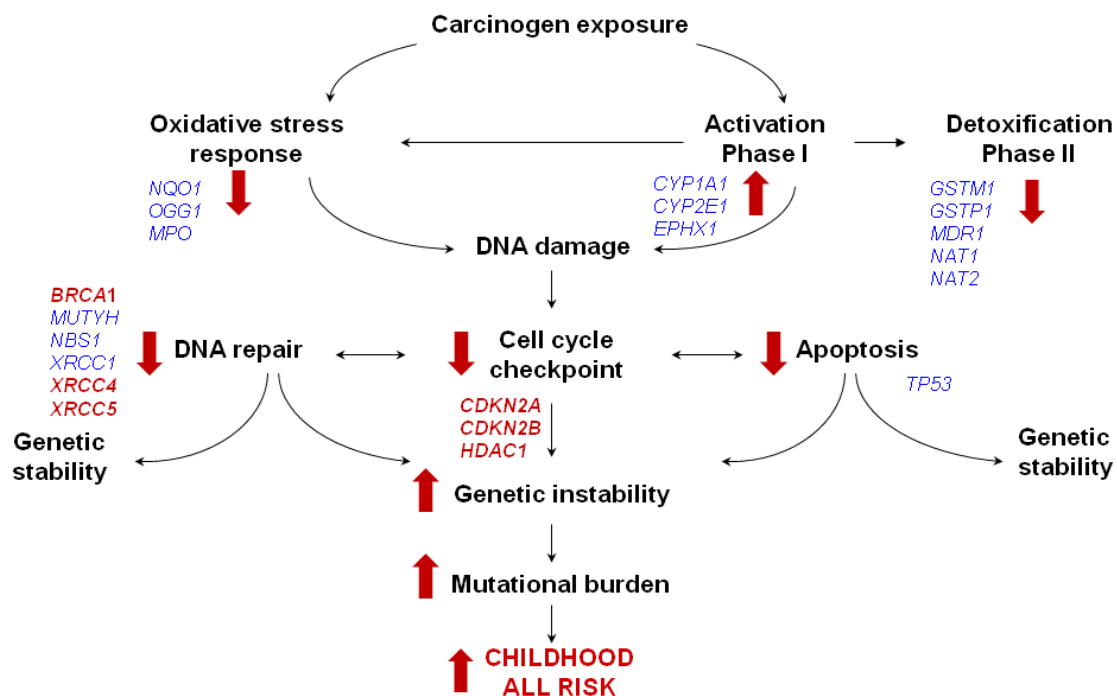


Figure 3. Cellular responses to environmental exposures and childhood ALL susceptibility

Biological processes known to contribute to cancer susceptibility include those involved in xenobiotic metabolism, oxidative stress response, DNA repair, cell cycle control and apoptosis. Candidate gene association studies have identified genes in many of these processes that are associated with ALL risk among children; shown in red are the genes identified through my doctoral work. Red arrows indicate how variants in these genes can influence their respective pathways and contribute to increased genetic instability, increased mutational burden and ultimately lead to childhood ALL.

Parental Genetics in Early-Onset Disorders

The early age of onset and evidence for the *in utero* origins of the disease support a role for parentally-mediated effects involved in the development of ALL in children. Interestingly, preferential loss of the maternally-derived 9p21 alleles has previously been demonstrated in leukemic cells further suggesting that a germline event involving 9p21 genes may play a role in the onset of childhood ALL (53). In Chapter Four, I investigated the potential role the mother's genes can play in shaping disease susceptibility in her offspring. The mother's genotype can influence gene expression in the fetus or directly perturb the intrauterine environment or even interact with the child's genes to modulate the overall risk of disease. It is difficult however to distinguish child from maternal genotype associations due to simple Mendelian inheritance. A susceptibility variant that is acting via the mother and that is therefore more prevalent among case-mothers as opposed to control-mothers, will be enriched among cases simply due to random transmission of maternal alleles to the offspring and can be mistook for an inherited effect within the child (54). In my study I attempted to dissect child and maternal genotype associations in childhood ALL using a hybrid case-parent trio/case-control design (55).

From a methodological point of view this work was imperative as existing methods for the detection of fetomaternal effects are tailored for specific study designs (e.g. case-trios only or case-mother/control-mother dyads). It was important for us, and potentially for other research groups simultaneously ascertaining population- and family-based datasets, to utilize all available genotypes in order to maximize analytical power in our study. However the adaptability and validity of available analytical methods for dissecting fetomaternal genotype associations using such a hybrid design had never been tested. And importantly the robustness of these methods, when parents of controls are missing and assumptions such as that of mating symmetry are

violated, had never been addressed. Mating asymmetry (MA) refers to the non-random distribution of alleles between males and females mating in the population. MA is to fetomaternal association testing what population stratification is to case-control association testing, that is an important source for confounding. In our study, we used simulations to show that the log-linear, likelihood-based framework of Weinberg and Umbach (56) using a case-triad/case-control design retains the ability to control for bias due to MA and can provide valid tests for maternal genetic effects even under MA, while other methods perform poorly in the presence of MA with considerable inflation of type I error rates. This study will aid in maximizing analytical efficiency to account for the underlying genetic complexities of early-onset disorders.

For the first time I showed that maternally-mediated genotype effects acting in combination with the child's genotype via cell cycle inhibitor genes *CDKN2A* (rs36228834) and *CDKN2B* (rs36229158) can modulate risk of childhood ALL. These studies complement our initial results which showed that inherited susceptibility to childhood ALL could be mediated by loss of cell cycle inhibition in the hematopoietic cells in the bone marrow of the child. In addition, reduced expression of *CDKN2A* and *CDKN2B* in the mother might also contribute to disease pathogenesis in her offspring. A possible explanation for this could be via changes in placental structure and function and disruption of the maternal-fetal interface. In fact reduced maternal CDKI expression could affect the size of the placenta and affect nutrient and hormonal transport to the fetus and perhaps provide the link between childhood ALL and high birth weight. At a crucial time in development when rapid growth and developmental changes are occurring and cell division rates are extremely high, tight regulation of gene expression is crucial. Variation in gene dosage of critical cell cycle control genes in both mother and child could cause increased proliferative stress on the growing fetus and lead to increased genomic instability and potentially contribute to leukemogenesis.

Though not yet published, I also identified a maternal genotype effect in double-strand break DNA repair gene *XRCC6* (rs132770) and a potential joint fetomaternal effect at loci *XRCC5* (rs11685387) (see Appendix III for details). *XRCC5* and *XRCC6* form a heterodimer that binds DNA double-strand breaks and facilitates repair by the nonhomologous end joining pathway (57). Low expression of the *XRCC5* and *XRCC6* genes leads to genomic instability and tumorigenesis (58). If activity of the *XRCC5* and *XRCC6* genes is inadequate in mothers then perturbations in the maternal environment could lead to increased mutational burden in the placenta and alter the function of placental proteins in key signaling and metabolic pathways and perhaps alter the intrauterine environment and contribute to disease susceptibility in the fetus. Furthermore, cytoprotective defense mechanisms are considerably less developed in the growing fetus therefore variation in genes involved in core processes such as maintenance of normal cell division and genomic integrity could have more profound consequences during fetal life (59).

The developmental origins of disease hypothesis, also called the Barker hypothesis, states that adverse influences early in development, and particularly during fetal life, can permanently change the child's physiology and metabolism in ways that can lead to increased disease risk later on in life (60). Well-established links between low birth weight and increased risk of coronary heart disease, diabetes, hypertension, and stroke in adulthood support the hypothesis for the *in utero* origins of disease (61) and there is also growing evidence that the intrauterine environment and enhanced prenatal exposure to estrogens, contributes to the predisposition of women to breast cancer in adulthood (62, 63). Though the mechanisms underlying the developmental origins of childhood ALL remain to be fully established, my study further corroborates the importance of investigating *in utero* genetic events to improve our understanding of the pathogenesis of childhood ALL. Additional investigation into other parentally-mediated genetic mechanisms such as maternal-fetal epistasis and parent-of-

origin effects are also required in order to fully understand genetic susceptibility to this early-onset disorder.

Limitations of the Study

The most obvious limitation of my study is power (see Appendix II for detailed power analysis); a total of 321 cases (203 family trios) and 329 controls were genotyped. Though these figures are far from those encountered in large-scale GWA studies, we have one of the largest childhood ALL trio samples worldwide. Availability of parental genotype data allowed us to investigate the role of maternal genetic effects in the susceptibility to childhood ALL and to show, for the first time, that the mother's genotype can influence the risk of leukemia in her offspring, further corroborating the importance of parental genetic contributions to early-onset disorders such as childhood leukemia. Due to the limited power of the study, it is possible that certain associations were missed and that effect sizes (odds ratios) were inflated and less precise (i.e. wider confidence intervals). Therefore not only may we have missed certain low-risk alleles, it is also possible that we have produced false-positive results due to over-estimation of the magnitude of the associations. On the other hand, it is encouraging to see that associated loci are consistent throughout the study and across methods. Still, results should be interpreted with caution and replication of the reported associations in larger, independent datasets remains crucial. There are also limitations associated with the statistical analysis. Given the limited power of the study, it was highly unlikely that we would be able to detect gene-gene interactions in my dataset, and subtype analysis was further constrained by even smaller sample sizes. Multivariate analysis including several SNPs and adjusting for potential confounders could also produce unreliable results due to over-parameterization.

While it is obviously more appealing to have a large, well-powered study sample this isn't always feasible. I argue that results from a well-designed small study, if interpreted carefully and if not overstated, could yield valuable insight into the genetic underpinnings of a disease. Small studies could provide a good setting for exploratory analyses, such as our investigation of fetomaternal effects, and could be used as the basis to design larger confirmatory studies.

It is also important at this time to discuss the limited power of our study with regard to the GWAS replication (Chapter Two, "Replication analysis confirms the association of *ARID5B* with childhood B-cell acute lymphoblastic leukemia", p. 167). In this study we successfully replicated five of the fifteen loci interrogated (annotating *ARID5B*) however we did not replicate two other strong GWAS signals (particularly *IKZF1* which was the strongest signal in one of the initial genome-wide studies (64), as well as *DDC* and *CEBPE*). The most likely culprit for non-replication is insufficient power; we had 80% power at the 5% level, to detect a minimum OR of 1.8 with RAFs $\geq 20\%$ and of 2.1 with RAFs $\geq 10\%$. Initial risk estimates were already below 1.8 and due to the "winner's curse" effect, which stipulates that the odds ratio of a disease variant is usually overestimated in the study that first describes it (65), it is possible that the true effect sizes at these loci may be too low for us to detect in our study sample. Lack of replication could also partially reflect between-study heterogeneity in terms of phenotypic classification of B-cell ALL patients, population-specific differences (though risk allele frequencies in cases did not differ between study cohorts) or heterogeneity in exposure to environmental factors (66).

Another interesting possibility is that the causal variant at *IKZF1* interacts with one or more other functional polymorphisms to influence disease and that these interacting loci have different allele frequencies between study populations. Greene et al. showed that the power to replicate statistically significant

independent main effects of a polymorphism can drop dramatically with a change of allele frequency of less than 0.1 at a second interacting polymorphism (67). Therefore if the true genetic model is epistatic, attempting to replicate a main effect is likely to fail unless allele frequencies of the interacting loci are close to identical. Study populations were comparable between studies (i.e. of European descent) but were not identical. Cryptic differences in the genetic structure of the French-Canadian “founder” population, compared to the US and UK populations used in the initial GWAS, could account for the lack of replication at these loci. Failure to replicate in our cohort does not necessarily invalidate the original findings but further validates the implication of *ARID5B* in childhood ALL susceptibility. And on the other hand, until additional replication studies with larger case-control samples are performed, we cannot rule out the possibility that the initial GWAS findings were false-positives.

FUTURE PERSPECTIVES

Integrative Genomics

The challenge in dissecting the genetic architecture of childhood ALL lies in its inherent complexity and in this study I have shown that insight into the true nature of childhood leukemia can only be gained by combining multiple analytical approaches and study designs. However, a complete understanding of the signalling pathways involved in leukemogenesis will likely only be achieved through systems-level approaches and the integration of diverse data types such as genomics, transcriptomics and proteomics data (68). Therefore, to completely embrace the complexity of childhood ALL, and of cancer in general, a goal should be to develop more comprehensive analytical approaches to integrate various different data sources rather than continue to use oversimplified one-dimensional methods of analysis (69). Future analyses

integrating different sources of 'omic' data for childhood ALL including genomic SNP (64, 70, 71) and copy number alteration arrays (72, 73), comparative genomic hybridization arrays (74, 75), molecular cytogenetic analyses (76), global gene expression profiles (49, 77), genome-wide methylation profiles (78, 79), and microRNA expression profiles (80), will allow a more comprehensive understanding of the mechanisms leading to the pathogenesis of childhood ALL and of its various subtypes. What is certain is that as we move forward into the post-genomic era, integration of large-scale datasets will be crucial to better understand the genetic basis of childhood ALL and the molecular mechanisms driving leukemogenesis.

Gene-environment interactions

The environment plays an important role in modulating interindividual variability in disease risk and while my doctoral research focused primarily on the genetic basis of childhood ALL, it is likely that an important etiologic component of this disease is exposure to environmental factors either acting alone or through complex gene-environment interactions involving both mother and child, and potentially also the father. For example, the distinct patterns of childhood acute lymphoblastic leukemia observed within race, gender, age and socioeconomic groups could indicate race- or gender-specific variability in genetic susceptibility, or could indicate differential exposure to as yet unidentified environmental factors. Or it could be that exposure to particular environmental agents during a vulnerable period in development (i.e. fetal life), increases cancer risk in genetically predisposed subgroups through gene-environment interactions. Given the mounting evidence that genetic variability, particularly in genes whose products metabolize carcinogens, repair DNA damage, and control the cell cycle, can influence an individual's susceptibility (or resistance) to cancer by modulating their response to the environment (81), investigation into the role of exogenous (e.g. pesticides, radiation) and endogenous factors (e.g. hormones),

and of their interaction with genes such as those studied here to influence childhood ALL susceptibility is warranted. The integration of environmental factors in genetic association studies can be extremely useful for further understanding the biological mechanisms involved in disease and gene-environment studies may help illuminate additional genetic factors involved in childhood ALL and help identify modifiable environmental factors for effective disease prevention (82).

Candidate Gene Versus Genome-Wide Association Studies

An important question raised at the beginning of my thesis was whether there was still need for candidate gene approaches in the era of genomics. The success of the more targeted candidate gene/pathway approaches relies largely on the current knowledge of the disease. By choosing functionally relevant variants in biologically relevant genes, it may be easier to identify the true causal variants through candidate gene studies, however conversely, a more agnostic approach may yield novel insight into the mechanisms of disease pathogenesis that otherwise may never have been investigated. For example the link between autophagy, a biological process that mediates degradation of cytoplasmic components in lysosomes and vacuoles, and Crohn's Disease was made evident through genome-wide association studies (83-85). The ideal approach to identifying genetic susceptibility loci might involve a combination of both philosophies. Methods for marrying hypothesis-driven approaches with whole-genome studies have been proposed, for example use of pathway knowledge to inform the analysis of whole-genome data and data mining procedures to detect patterns of interaction and novel pathways implicated in disease have been suggested (82). Another approach involves creating gene-centric and functionally relevant genome-wide SNP arrays for GWAS (86). I argue however that candidate gene/pathway approaches can be a useful first step in exploring potential causal pathways and biologic mechanisms involved in the pathogenesis of complex diseases with multifactorial inheritance such as

early-onset disorders. Elucidation of the genetic mechanisms through which genes act to influence disease risk, e.g. through parentally-mediated genetic effects, may better be depicted through focused candidate gene association studies. What is certain is that to reach the goal of dissecting the genetic architecture of childhood ALL, a combination of many different approaches and study designs will be required including both candidate gene/pathway and genome-wide association studies.

Common Versus Rare Variants

The common disease, common variant (CDCV) hypothesis has largely dictated how we search for genetic susceptibility loci in complex diseases. In addition to the many candidate gene studies that have been published over the years, most if not all of the novel genetic susceptibility loci identified through GWAS pertain to common genetic variants with minor allele frequencies typically exceeding 5%, the effect sizes of which are small, explaining but a finite portion of the genetic variability associated with human diseases (87-89). This is because whole-genome studies rely on a complex disease model in which common DNA sequence variants with widespread, but marginal, effects predominate; however it may not be that surprising that common variants provide little help in predicting complex disease risk. The arguable lack of success of GWAS in explaining the heritable component of complex disease susceptibility (90, 91) has led to a recent paradigm shift in genetic association testing toward rare variants, with presumed strong effects, to explain the missing heritability in complex traits (Figure 4). There is strong evidence that rare variants are involved in complex disease etiology (92-95) yet the multiple rare variant (MRV) paradigm remains unexplored in childhood ALL. With the advent of high-throughput next-generation sequencing technologies, there is a current push towards deep resequencing of entire exomes, and as costs continue to decrease whole genomes, in large cohorts to detect the full catalogue of genetic

variation (common and rare, sequence and structural) associated with complex traits. And it may well be that full appreciation of the importance of rare genetic variants in complex disease can only be captured through pathway-based analysis strategies, given that the combination of rare sequence variants from multiple genes functioning in the same biological pathway is likely to be more relevant than the individual effects of single rare sequence variants. Next-generation sequencing promises to provide an unprecedented view of the genetic landscape of complex diseases and of the genetic basis of interindividual variability in the susceptibility to childhood ALL (96-98) but in order for this genomic data to have a meaningful impact on our understanding of the disease, I maintain that an effort must be made to integrate sequencing data with all the additional 'omic' information that is being generated.

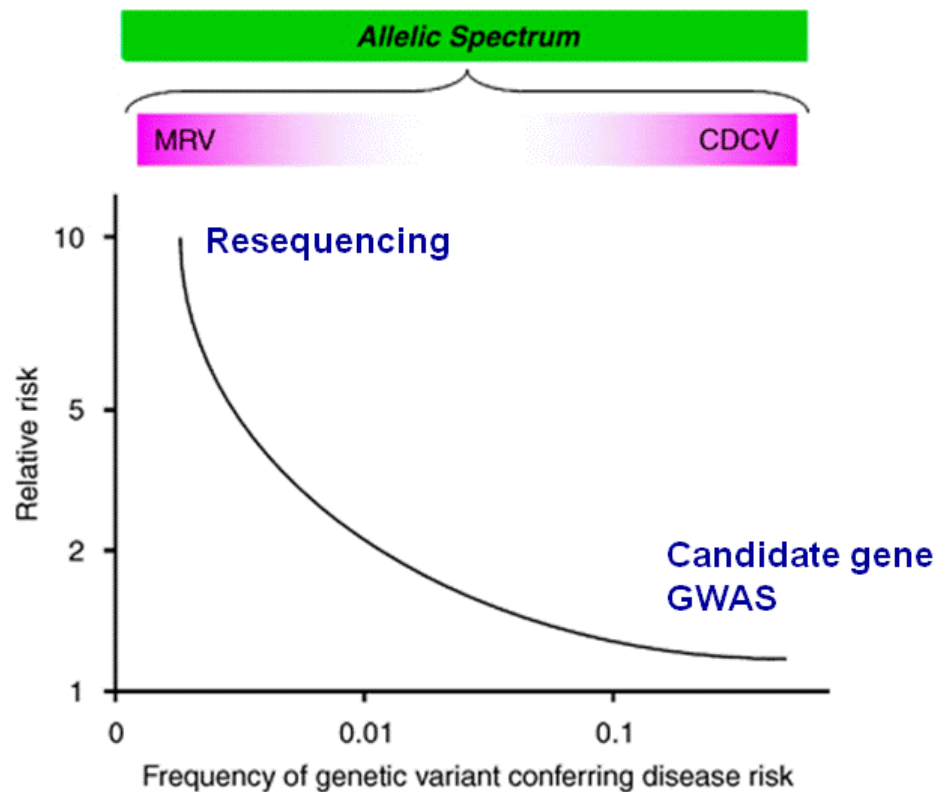


Figure 4. A paradigm shift in genetic association testing – from common to rare variants

The genetic architecture of complex traits likely involves a wide allelic spectrum ranging from rare to common genetic variants. The common disease, common variant (CDCV) model postulates the existence of genetic variants that are common and that confer modest to low effects on disease risk, and has long dictated candidate gene and genome-wide association studies (GWAS). The multiple rare variant (MRV) model holds that complex traits result from many different mutations each of which is individually rare but with very strong combined effects, and is at the basis of the emergence of next-generation resequencing technologies. These models are not exclusive and it is likely that both contribute to interindividual variability in complex disease risk.

Figure adapted from the Psychiatric GWAS Consortium Steering Committee, 2009 (99).

CONCLUSION

While established genetic risk factors able to predict ALL risk among children are still lacking, this study will help point towards new avenues of research and provides a strong rationale for further elucidation of the mechanisms through which aberrant cell cycle control and double-strand break repair are involved in childhood ALL susceptibility. The results of my study also serve to illustrate the complexity of childhood cancer and the need for a variety of methods and study designs able to capture the underlying genetic intricacies involved in disease pathogenesis. Despite the recent shift toward rare SNPs, expected to be more functionally relevant than common SNPs in disease susceptibility (88, 95), a more realistic disease model is that functional variants involved in ALL have a wide spectrum of allele frequencies that range from common to not-so-common to rare and that the landscape of genetic variation contributing to ALL predisposition includes structural variation, epigenetic changes, and even parentally-mediated genetic effects. Yet no matter the architecture, genetic susceptibility to ALL can only comprehensively be identified through the use and integration of multiple complementary analytical approaches.

Identification of genetic susceptibility factors along with a better understanding of the underlying mechanisms that drive leukemogenesis, will help refine risk classification and tailored disease management, offering new opportunities for personalized medicine. This is important given that, despite improved survival rates in childhood cancer, treatment is not optimal and ALL is still the leading cause of death among children. Another important point to consider is the late effects of therapy. Currently, approximately one in every 400 young adult is a childhood cancer survivor (100). Adult survivors of childhood ALL represent a new and growing population that did not exist just a few decades ago, a population that is at increased risk for chronic health conditions including

obesity, hypertension, type II diabetes, depression, as well as secondary malignancies due to late-occurring effects of their treatment (101-103). The genetic susceptibility loci identified in this study could lead to new therapeutic targets for childhood ALL and help reduce the debilitating side effects related to the cancer and its treatment.

Ultimately, it is my hope that this study will help shed light on the genetic architecture of ALL and thereby contribute to better diagnosis of the patients, to more refined treatment protocols, and hopefully lead to overall prevention of the disease in children. Though it seems increasingly unlikely that genetic prediction will ever be accurate enough for complex multifactorial diseases, the identification of established genetic, and eventually environmental, susceptibility factors in childhood ALL will certainly be useful in identifying at-risk groups and establishing new public health policies to guide lifestyle and health-related behavioral changes. This research will provide greater insight into the etiologic intricacies of childhood ALL and will bring us one step closer to that ultimate goal of decreasing mortality by reducing risk and improving diagnosis and treatment.

REFERENCES

1. Sherborne AL, Hosking FJ, Prasad RB, Kumar R, Koehler R, Vijayakrishnan J, et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. *Nat Genet.* 2010 May 9.
2. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet.* 2007 May;39(5):645-9.
3. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science.* 2007 Jun 8;316(5830):1491-3.
4. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, et al. A common allele on chromosome 9 associated with coronary heart disease. *Science.* 2007 Jun 8;316(5830):1488-91.
5. Gschwendtner A, Bevan S, Cole JW, Plourde A, Matarin M, Ross-Adams H, et al. Sequence variants on chromosome 9p21.3 confer risk for atherosclerotic stroke. *Ann Neurol.* 2009 May;65(5):531-9.
6. Helgadottir A, Thorleifsson G, Magnusson KP, Gretarsdottir S, Steinthorsdottir V, Manolescu A, et al. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nat Genet.* 2008 Feb;40(2):217-24.
7. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science.* 2007 Jun 1;316(5829):1336-41.
8. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science.* 2007 Jun 1;316(5829):1341-5.
9. Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B, et al. Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet.* 2009 Aug;41(8):899-904.
10. Wrensch M, Jenkins RB, Chang JS, Yeh RF, Xiao Y, Decker PA, et al. Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nat Genet.* 2009 Aug;41(8):905-8.
11. Bishop DT, Demenais F, Iles MM, Harland M, Taylor JC, Corda E, et al. Genome-wide association study identifies three loci associated with melanoma risk. *Nat Genet.* 2009 Aug;41(8):920-5.

12. Debniak T, Gorski B, Huzarski T, Byrski T, Cybulski C, Mackiewicz A, et al. A common variant of CDKN2A (p16) predisposes to breast cancer. *J Med Genet.* 2005 Oct;42(10):763-5.
13. Driver KE, Song H, Lesueur F, Ahmed S, Barbosa-Morais NL, Tyrer JP, et al. Association of single-nucleotide polymorphisms in the cell cycle genes with breast cancer in the British population. *Carcinogenesis.* 2008 Feb;29(2):333-41.
14. Gayther SA, Song H, Ramus SJ, Kjaer SK, Whittemore AS, Quaye L, et al. Tagging single nucleotide polymorphisms in cell cycle control genes and susceptibility to invasive epithelial ovarian cancer. *Cancer Res.* 2007 Apr 1;67(7):3027-35.
15. Chen J, Li D, Wei C, Sen S, Killary AM, Amos CI, et al. Aurora-A and p16 polymorphisms contribute to an earlier age at diagnosis of pancreatic cancer in Caucasians. *Clin Cancer Res.* 2007 May 15;13(10):3100-4.
16. Kumar R, Smeds J, Berggren P, Straume O, Rozell BL, Akslen LA, et al. A single nucleotide polymorphism in the 3'untranslated region of the CDKN2A gene is common in sporadic primary melanomas but mutations in the CDKN2B, CDKN2C, CDK4 and p53 genes are rare. *Int J Cancer.* 2001 Nov 20;95(6):388-93.
17. Healy J, Belanger H, Beaulieu P, Lariviere M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood.* 2007 Jan 15;109(2):683-92.
18. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* 2000 Jan 7;100(1):57-70.
19. Levine AJ. p53, the cellular gatekeeper for growth and division. *Cell.* 1997 Feb 7;88(3):323-31.
20. Sherr CJ, McCormick F. The RB and p53 pathways in cancer. *Cancer Cell.* 2002 Aug;2(2):103-12.
21. Bertin R, Acquaviva C, Mirebeau D, Guidal-Giroux C, Vilmer E, Cave H. CDKN2A, CDKN2B, and MTAP gene dosage permits precise characterization of mono- and bi-allelic 9p21 deletions in childhood acute lymphoblastic leukemia. *Genes Chromosomes Cancer.* 2003 May;37(1):44-57.
22. Novara F, Beri S, Bernardo ME, Bellazzi R, Malovini A, Ciccone R, et al. Different molecular mechanisms causing 9p21 deletions in acute lymphoblastic leukemia of childhood. *Hum Genet.* 2009 Oct;126(4):511-20.
23. Guo SX, Taki T, Ohnishi H, Piao HY, Tabuchi K, Bessho F, et al. Hypermethylation of p16 and p15 genes and RB protein expression in acute leukemia. *Leuk Res.* 2000 Jan;24(1):39-46.

24. Cunnington MS, Santibanez Koref M, Mayosi BM, Burn J, Keavney B. Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression. *PLoS Genet.* 2010;6(4):e1000899.
25. Wu KH, Wang CH, Yang YL, Peng CT, Lin WD, Tsai FJ, et al. Significant association of XRCC4 single nucleotide polymorphisms with childhood leukemia in Taiwan. *Anticancer Res.* 2010 Feb;30(2):529-33.
26. Budman J, Kim SA, Chu G. Processing of DNA for nonhomologous end-joining is controlled by kinase activity and XRCC4/ligase IV. *J Biol Chem.* 2007 Apr 20;282(16):11950-9.
27. Chang CH, Chang CL, Tsai CW, Wu HC, Chiu CF, Wang RF, et al. Significant association of an XRCC4 single nucleotide polymorphism with bladder cancer susceptibility in Taiwan. *Anticancer Res.* 2009 May;29(5):1777-82.
28. Chiu CF, Tsai MH, Tseng HC, Wang CL, Wang CH, Wu CN, et al. A novel single nucleotide polymorphism in XRCC4 gene is associated with oral cancer susceptibility in Taiwanese patients. *Oral Oncol.* 2008 Sep;44(9):898-902.
29. Chiu CF, Wang CH, Wang CL, Lin CC, Hsu NY, Weng JR, et al. A novel single nucleotide polymorphism in XRCC4 gene is associated with gastric cancer susceptibility in Taiwan. *Ann Surg Oncol.* 2008 Feb;15(2):514-8.
30. Tseng HC, Tsai MH, Chiu CF, Wang CH, Chang NW, Huang CY, et al. Association of XRCC4 codon 247 polymorphism with oral cancer susceptibility in Taiwan. *Anticancer Res.* 2008 May-Jun;28(3A):1687-91.
31. Hsu NY, Wang HC, Wang CH, Chang CL, Chiu CF, Lee HZ, et al. Lung cancer susceptibility and genetic polymorphism of DNA repair gene XRCC4 in Taiwan. *Cancer Biomark.* 2009;5(4):159-65.
32. Wang S, Wang M, Yin S, Fu G, Li C, Chen R, et al. A novel variable number of tandem repeats (VNTR) polymorphism containing Sp1 binding elements in the promoter of XRCC5 is a risk factor for human bladder cancer. *Mutat Res.* 2008 Feb 1;638(1-2):26-36.
33. Hayden PJ, Tewari P, Morris DW, Staines A, Crowley D, Nieters A, et al. Variation in DNA repair genes XRCC3, XRCC4, XRCC5 and susceptibility to myeloma. *Hum Mol Genet.* 2007 Dec 15;16(24):3117-27.
34. Liu Y, Zhang H, Zhou K, Chen L, Xu Z, Zhong Y, et al. Tagging SNPs in non-homologous end-joining pathway genes and risk of glioma. *Carcinogenesis.* 2007 Sep;28(9):1906-13.
35. Park JY, Huang Y, Sellers TA. Single nucleotide polymorphisms in DNA repair genes and prostate cancer risk. *Methods Mol Biol.* 2009;471:361-85.

36. Verkaik NS, Esveldt-van Lange RE, van Heemst D, Bruggenwirth HT, Hoeijmakers JH, Zdzienicka MZ, et al. Different types of V(D)J recombination and end-joining defects in DNA double-strand break repair mutant mammalian cells. *Eur J Immunol*. 2002 Mar;32(3):701-9.
37. Lewis SM, Wu GE. The origins of V(D)J recombination. *Cell*. 1997 Jan 24;88(2):159-62.
38. Oltz EM. Regulation of antigen receptor gene assembly in lymphocytes. *Immunol Res*. 2001;23(2-3):121-33.
39. Gao Y, Ferguson DO, Xie W, Manis JP, Sekiguchi J, Frank KM, et al. Interplay of p53 and DNA-repair protein XRCC4 in tumorigenesis, genomic stability and development. *Nature*. 2000 Apr 20;404(6780):897-900.
40. Wiemels JL, Leonard BC, Wang Y, Segal MR, Hunger SP, Smith MT, et al. Site-specific translocation and evidence of postnatal origin of the t(1;19) E2A-PBX1 fusion in childhood acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A*. 2002 Nov 12;99(23):15101-6.
41. Greaves MF, Wiemels J. Origins of chromosome translocations in childhood leukaemia. *Nat Rev Cancer*. 2003 Sep;3(9):639-49.
42. Wang Y, Cortez D, Yazdi P, Neff N, Elledge SJ, Qin J. BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes Dev*. 2000 Apr 15;14(8):927-39.
43. Scardocci A, Guidi F, D'Alo F, Gumiero D, Fabiani E, Diruscio A, et al. Reduced BRCA1 expression due to promoter hypermethylation in therapy-related acute myeloid leukaemia. *Br J Cancer*. 2006 Oct 23;95(8):1108-13.
44. Friedenson B. The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers. *BMC Cancer*. 2007;7:152.
45. Wolffe AP. Histone deacetylase: a regulator of transcription. *Science*. 1996 Apr 19;272(5260):371-2.
46. Magnaghi-Jaulin L, Groisman R, Naguibneva I, Robin P, Lorain S, Le Villain JP, et al. Retinoblastoma protein represses transcription by recruiting a histone deacetylase. *Nature*. 1998 Feb 5;391(6667):601-5.
47. Wada T, Kikuchi J, Nishimura N, Shimizu R, Kitamura T, Furukawa Y. Expression levels of histone deacetylases determine the cell fate of hematopoietic progenitors. *J Biol Chem*. 2009 Oct 30;284(44):30673-83.
48. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*. 2002 Mar;1(2):133-43.

49. Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, Williams WK, et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*. 2003 Oct 15;102(8):2951-9.
50. Bonifer C, Bowen DT. Epigenetic mechanisms regulating normal and malignant haematopoiesis: new therapeutic targets for clinical medicine. *Expert Rev Mol Med*. 2010;12:e6.
51. Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P, et al. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiol Biomarkers Prev*. 2004 Jun;13(6):1013-21.
52. Hung RJ, Baragatti M, Thomas D, McKay J, Szeszenia-Dabrowska N, Zaridze D, et al. Inherited predisposition of lung cancer: a hierarchical modeling approach to DNA repair and cell cycle control pathways. *Cancer Epidemiol Biomarkers Prev*. 2007 Dec;16(12):2736-44.
53. Morison IM, Ellis LM, Teague LR, Reeve AE. Preferential loss of maternal 9p alleles in childhood acute lymphoblastic leukemia. *Blood*. 2002 Jan 1;99(1):375-7.
54. Buyske S. Maternal genotype effects can alias case genotype effects in case-control studies. *Eur J Hum Genet*. 2008 Jul;16(7):783-5.
55. Healy J, Bourgey M, Richer C, Sinnett D, Roy-Gagnon M-H. Detection of Fetomaternal Genotype Associations in Early-Onset Disorders: Evaluation of Different Methods and Their Application to Childhood Leukemia. *Journal of Biomedicine and Biotechnology*. 2010(doi:10.1155/2010/369534):13.
56. Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am J Hum Genet*. 2005 Oct;77(4):627-36.
57. Pastwa E, Blasiak J. Non-homologous DNA end joining. *Acta Biochim Pol*. 2003;50(4):891-908.
58. Gullo C, Au M, Feng G, Teoh G. The biology of Ku and its potential oncogenic role in cancer. *Biochim Biophys Acta*. 2006 Apr;1765(2):223-34.
59. Wan J, Winn LM. In utero-initiated cancer: the role of reactive oxygen species. *Birth Defects Res C Embryo Today*. 2006 Dec;78(4):326-32.
60. Barker DJ. The origins of the developmental origins theory. *J Intern Med*. 2007 May;261(5):412-7.
61. de Boo HA, Harding JE. The developmental origins of adult disease (Barker) hypothesis. *Aust N Z J Obstet Gynaecol*. 2006 Feb;46(1):4-14.

62. Xue F, Michels KB. Intrauterine factors and risk of breast cancer: a systematic review and meta-analysis of current evidence. *Lancet Oncol.* 2007 Dec;8(12):1088-100.
63. Soto AM, Vandenberg LN, Maffini MV, Sonnenschein C. Does breast cancer start in the womb? *Basic Clin Pharmacol Toxicol.* 2008 Feb;102(2):125-33.
64. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, Price A, Olver B, Sheridan E, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet.* 2009 Sep;41(9):1006-10.
65. Zollner S, Pritchard JK. Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am J Hum Genet.* 2007 Apr;80(4):605-15.
66. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, et al. Replicating genotype-phenotype associations. *Nature.* 2007 Jun 7;447(7145):655-60.
67. Greene CS, Penrod NM, Williams SM, Moore JH. Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS One.* 2009;4(6):e5639.
68. Kitano H. Systems biology: a brief overview. *Science.* 2002 Mar 1;295(5560):1662-4.
69. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.* 2008 Oct 23;455(7216):1061-8.
70. Trevino LR, Yang W, French D, Hunger SP, Carroll WL, Devidas M, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet.* 2009 Sep;41(9):1001-5.
71. Han S, Lee KM, Park SK, Lee JE, Ahn HS, Shin HY, et al. Genome-wide association study of childhood acute lymphoblastic leukemia in Korea. *Leuk Res.* 2010 Feb 26.
72. Bateman CM, Colman SM, Chaplin T, Young BD, Eden TO, Bhakta M, et al. Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia. *Blood.* 2010 Apr 29;115(17):3553-8.
73. Emerenciano M, Bungaro S, Cazzaniga G, Dorea MD, Coser VM, Magalhaes IQ, et al. ETV6-RUNX1 fusion gene and additional genetic changes in infant leukemia: a genome-wide analysis. *Cancer Genet Cytogenet.* 2009 Sep;193(2):86-92.
74. Larramendy ML, Huhta T, Vettenranta K, El-Rifai W, Lundin J, Pakkala S, et al. Comparative genomic hybridization in childhood acute lymphoblastic leukemia. *Leukemia.* 1998 Oct;12(10):1638-44.

75. Tsuzuki S, Karnan S, Horibe K, Matsumoto K, Kato K, Inukai T, et al. Genetic abnormalities involved in t(12;21) TEL-AML1 acute lymphoblastic leukemia: analysis by means of array-based comparative genomic hybridization. *Cancer Sci.* 2007 May;98(5):698-706.
76. Konn ZJ, Wright SL, Barber KE, Harrison CJ. Fluorescence In situ hybridization (FISH) as a tool for the detection of significant chromosomal abnormalities in childhood leukaemia. *Methods Mol Biol.* 2009;538:29-55.
77. Den Boer ML, van Slegtenhorst M, De Menezes RX, Cheok MH, Buijs-Gladdines JG, Peters ST, et al. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *Lancet Oncol.* 2009 Feb;10(2):125-34.
78. Davidsson J, Lilljebjorn H, Andersson A, Veerla S, Heldrup J, Behrendtz M, et al. The DNA methylome of pediatric acute lymphoblastic leukemia. *Hum Mol Genet.* 2009 Nov 1;18(21):4054-65.
79. Dunwell T, Hesson L, Rauch TA, Wang L, Clark RE, Dallol A, et al. A genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers. *Mol Cancer.* 2010;9:44.
80. Ju X, Li D, Shi Q, Hou H, Sun N, Shen B. Differential microRNA expression in childhood B-cell precursor acute lymphoblastic leukemia. *Pediatr Hematol Oncol.* 2009 Jan;26(1):1-10.
81. Rothman N, Wacholder S, Caporaso NE, Garcia-Closas M, Buetow K, Fraumeni JF, Jr. The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. *Biochim Biophys Acta.* 2001;1471(2):C1-10.
82. Thomas D. Gene-environment-wide association studies: emerging approaches. *Nat Rev Genet.* 2010 Mar 9;11(4):259-72.
83. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet.* 2007 May;39(5):596-604.
84. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007 Jun 7;447(7145):661-78.
85. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat Genet.* 2007 Feb;39(2):207-11.
86. Jorgenson E, Witte JS. A gene-centric approach to genome-wide association studies. *Nat Rev Genet.* 2006 Nov;7(11):885-91.

87. Kraft P, Hunter DJ. Genetic risk prediction--are we there yet? *N Engl J Med*. 2009 Apr 23;360(17):1701-3.
88. Goldstein DB. Common genetic variation and human traits. *N Engl J Med*. 2009 Apr 23;360(17):1696-8.
89. Dermitzakis ET, Clark AG. Genetics. Life after GWA studies. *Science*. 2009 Oct 9;326(5950):239-40.
90. Hirschhorn JN. Genomewide association studies--illuminating biologic pathways. *N Engl J Med*. 2009 Apr 23;360(17):1699-701.
91. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol*. 2010;8(1):e1000294.
92. Azzopardi D, Dallosso AR, Eliason K, Hendrickson BC, Jones N, Rawstorne E, et al. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. *Cancer Res*. 2008 Jan 15;68(2):358-63.
93. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*. 2004 Aug 6;305(5685):869-72.
94. Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet*. 2008 May;40(5):592-9.
95. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet*. 2008 Jan;82(1):100-12.
96. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*. 2010 Jun;11(6):415-25.
97. Via M, Gignoux C, Burchard EG. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med*. 2010;2(1):3.
98. Mardis ER. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. *Genome Med*. 2009;1(4):40.
99. A framework for interpreting genome-wide association studies of psychiatric disorders. *Mol Psychiatry*. 2009 Jan;14(1):10-7.
100. Hawkins MM, Stevens MC. The long-term survivors. *Br Med Bull*. 1996 Oct;52(4):898-923.

101. Armstrong GT, Sklar CA, Hudson MM, Robison LL. Long-term health status among survivors of childhood cancer: does sex matter? *J Clin Oncol*. 2007 Oct 1;25(28):4477-89.
102. Nathan PC, Wasilewski-Masker K, Janzen LA. Long-term outcomes in survivors of childhood acute lymphoblastic leukemia. *Hematol Oncol Clin North Am*. 2009 Oct;23(5):1065-82, vi-vii.
103. Pui CH, Cheng C, Leung W, Rai SN, Rivera GK, Sandlund JT, et al. Extended follow-up of long-term survivors of childhood acute lymphoblastic leukemia. *N Engl J Med*. 2003 Aug 14;349(7):640-9.

APPENDIX I

MEASURES OF ASSOCIATION

In population-based association studies, an association between a genetic risk factor and disease outcome can be measured by directly comparing disease frequency differences between individuals with and without the risk factor, or by comparing the odds of disease.

Table I. Presentation of data from a cohort study or a case-control study of a binary risk factor

Disease	Risk factor		Total
	Present	Absent	
Present	a	b	$a + b$
Absent	c	d	$c + d$
Total	$a + c$	$b + d$	n

I = probability of disease among individuals with the risk factor = $a/(a + c)$.

U = probability of disease among individuals without the risk factor = $b/(b + d)$

P = probability of carrying the risk factor among cases = $a/(a + b)$.

Q = probability of carrying the risk factor among controls = $c/(c + d)$

The relative risk (RR) is the ratio of risks (probability of developing the disease) between carriers and non-carriers of the risk factor:

$$RR = \frac{I}{U} = \frac{a/(a+c)}{b/(b+d)}.$$

And the difference in risk, also known as the attributable risk is simply:

$$I - U = \frac{a}{a + c} - \frac{b}{b + d}.$$

Measures of relative risk can only be obtained from cohort designs, in which individuals exposed and not exposed to the risk are followed over a period of time to identify those that develop the disease.

The incidence refers to the rate at which previously unaffected individuals in a given population develop disease, and the prevalence can be defined as the proportion of individuals in a population who are affected; each are measures of disease frequency.

Because of sampling design, measures of risk cannot be directly measured in a case-control design in which individuals are ascertained based on disease status. A more useful measure of association is the odds ratio (OR) in which the odds of carrying the risk factor among affected individuals is compared to the odds of carrying the risk factor among unaffected individuals.

Odds of carrying the risk factor among cases: $P/(1 - P) = \frac{a/(a+b)}{(1-(a/(a+b)))}$

Odds of carrying the risk factor among controls: $Q/(1 - Q) = \frac{c/(c+d)}{(1-(c/(c+d)))}$

$$OR = \frac{P/(1 - P)}{Q/(1 - Q)} = \frac{\frac{a/(a+b)}{b/(a+b)}}{\frac{c/(c+d)}{d/(c+d)}} = \frac{ad}{bc}$$

The variance of the OR can be estimated by:

$$var(\ln OR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

And a 95% confidence interval (95%CI) by:

$$\exp[\ln OR \pm 1.96 \sqrt{var(\ln OR)}]$$

The OR of exposure in a case-control study is identical to the OR of disease in a cohort study and for a rare disease, the OR approximates the relative risk (1).

The OR is a measure of effect size, describing the strength of association between a risk factor and the disease. An OR = 1 if both the risk factor and disease outcome are independent, if OR>1 there is evidence of a positive effect of the risk factor on increased disease risk while OR<1 suggests a protective effect. The calculation for the OR can be modified to include multiple levels of risk exposure or linear measures of exposure and to accommodate the sampling procedure (e.g. pair-matched case-control study) (2).

Besides the 2 x 2 contingency table, another way of generating an OR beyond two binary variables is by using logistic regression (3). The effect of a risk factor, say a SNP, on disease outcome can be measured through logistic regression as follows:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

where p is the probability of disease occurring, $p/(1-p)$ the odds of that event occurring, X_1 is the risk factor (e.g. SNP alleles coded 0 or 1) and the coefficient of β_1 is the effect of the SNP on disease risk.

The OR is obtained through the following transformation:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 \quad \rightarrow \quad OR_{X^1} = \exp(\beta_1)$$

Using the data in Table I a test of the null hypothesis that the risk factor is not associated with disease, $H_0: OR = 1$, is given by the chi-square test:

$$X^2 = \frac{[Observed - Expected]^2}{Expected}$$

$$X^2 = \frac{[a - E(a)]^2}{E(a)} + \frac{[b - E(b)]^2}{E(b)} + \frac{[c - E(c)]^2}{E(c)} + \frac{[d - E(d)]^2}{E(d)}$$

Under the null hypothesis, X^2 has asymptotically (in large samples) a chi-square distribution with 1 degree of freedom (df).

For small sample sizes a more appropriate test for significance is the Fisher's exact test:

$$p = \frac{(a + b)!(c + d)!(a + c)!(b + d)!}{a!b!c!d!n!}$$

In family-based association designs consisting of case-trios (proband and both parents), the simplest association test is the transmission disequilibrium test (TDT) which compares observed allele distributions to those expected under Mendelian transmissions (4). Given that homozygous parents are uninformative, only heterozygote parents are used in the TDT. An advantage of family-based designs is that over-transmission of an allele to affected offspring is evidence of both linkage and association of the marker allele with a disease susceptibility locus.

Table II. Presentation of data from a family-based case-trio study of a binary risk factor

Non-transmitted allele	Transmitted allele		Total
	A	a	
A	<i>a</i>	<i>b</i>	<i>a + b</i>
a	<i>c</i>	<i>d</i>	<i>c + d</i>
Total	<i>a + c</i>	<i>b + d</i>	<i>n</i>

The appropriate matched analysis for the data presented in Table 1-2 is achieved using McNemar's chi-square test to see whether or not the a allele is transmitted more often than expected to affected offspring:

$$TDT = \frac{(b - c)^2}{(b + c)}$$

Under the null hypothesis of no linkage and no association, the TDT has asymptotically a chi-square distribution with 1 degree of freedom (df). Both linkage and association between the disease locus and the marker have to be present for the TDT to reject the null hypothesis.

The OR for the TDT can be estimated as, $OR = b/c$, with $var(\ln OR) = \frac{1}{b} + \frac{1}{c}$.

An extension of the TDT is the family-based association test (FBAT) which allows testing with trios with missing parents, haplotypes or markers with multiple alleles, different genetic models, different sampling designs, different disease phenotypes all in the same framework (5, 6).

Let X denote a variable that translates an offspring's genotype to a numeric value (for example X can count the number of a alleles, say 0 for AA, 1 for Aa and 2 for aa). Let P denote the genotype of the parents, and T is the coded offspring trait (say 0 for unaffected, 1 for affected). The covariance statistic used in the FBAT test is:

$$U = \sum T * (X - E(X|P))$$

where U is the covariance, $E(X|P)$ is the expected value of X computed under the null hypothesis, and summation is over all offspring in the sample. Mendel's laws of transmission underlie the calculation of $E(X|P)$.

The FBAT statistic is defined by dividing U^2 by its variance which is also calculated under the null hypothesis by conditioning on parental genotypes (see (7) for details):

$$Z = U/\sqrt{\text{var}(U)}, \text{ or equivalently, } X_{FBAT}^2 = U^2/\text{var}(U)$$

A positive Z score indicates over-transmission of the allele to affected offspring providing evidence of linkage and association with disease whereas a negative Z score indicates preferential under-transmission of the allele suggesting a protective effect.

For large samples, Z is approximately distributed $N(0,1)$ and X_{FBAT}^2 has asymptotically a chi-square distribution with 1 df. The FBAT statistic is exactly the same as the TDT statistic:

$$X_{FBAT}^2 = X_{TDT}^2$$

When both parents are genotyped, only affected offspring are included in the model, and an additive mode of inheritance is assumed, that is X counts the

number of alleles. Changing how T is defined allows us to include unaffected offspring or fit alternative or multiple traits, while changing how X is defined allows us to test alternative genetic models or incorporate multiple alleles (e.g. haplotypes) (5).

REFERENCES

1. Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. *Am J Epidemiol.* 1982 Sep;116(3):547-53.
2. Thomas DC. *Statistical Methods in Genetic Epidemiology*: Oxford University Press; 2004.
3. Breslow NE, Day NE. *Statistical methods in cancer research. Volum 1 - The analysis of case-control studies*: International Agency for Research on Cancer Lyon; 1980.
4. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet.* 1993 Mar;52(3):506-16.
5. Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered.* 2000 Jul-Aug;50(4):211-23.
6. Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. *Genet Epidemiol.* 2000;19 Suppl 1:S36-42.
7. Horvath S, Xu X, Laird NM. The family based association test method: strategies for studying general genotype--phenotype associations. *Eur J Hum Genet.* 2001 Apr;9(4):301-6.

APPENDIX II

*POWER OF THE ASSOCIATION STUDIES AND
COMPLETE RESULTS OF THE CASE-CONTROL
AND FAMILY-BASED ASSOCIATION TESTS*

POWER OF THE ASSOCIATION STUDY

Table I. Power calculations for main effects using a case-control design

D'	OR	P_M	Recessive			Dominant			Multiplicative		
			$P_D:P_M$	1:1.5	1:1	1.5:1	1:1.5	1:1	1.5:1	1:1.5	1:1
0.8	1.25	0.05	0.0500	0.0502	0.0504	0.0765	0.1071	0.1003	0.0787	0.1146	0.1106
		0.1	0.0503	0.0513	0.0526	0.0975	0.1453	0.1229	0.1062	0.1727	0.1569
		0.2	0.0518	0.0592	0.0654	0.1194	0.1776	0.1195	0.1479	0.2644	0.2070
		0.3	0.0554	0.0767	0.0852	0.1251	0.1719	0.0922	0.1776	0.3235	0.2083
		0.4	0.0609	0.1032	0.0984	0.1196	0.1475	0.0672	0.1939	0.3534	0.1731
	1.5	0.05	0.0501	0.0507	0.0515	0.1503	0.2593	0.2315	0.1604	0.2904	0.2737
		0.1	0.0511	0.0552	0.0623	0.2267	0.3835	0.3012	0.2641	0.4803	0.4252
		0.2	0.0572	0.0861	0.1094	0.2978	0.4649	0.2750	0.4077	0.6987	0.5524
		0.3	0.0712	0.1540	0.1806	0.3087	0.4344	0.1800	0.4948	0.7910	0.5391
		0.4	0.0925	0.2526	0.2194	0.2832	0.3534	0.1007	0.5348	0.8233	0.4305
	2	0.05	0.0505	0.0527	0.0558	0.3976	0.6578	0.5849	0.4374	0.7305	0.6926
		0.1	0.0542	0.0704	0.0898	0.5940	0.8335	0.6956	0.6950	0.9328	0.8840
		0.2	0.0783	0.1902	0.2693	0.7119	0.8868	0.6043	0.8855	0.9925	0.9494
		0.3	0.1327	0.4254	0.4723	0.7071	0.8385	0.3750	0.9399	0.9978	0.9293
		0.4	0.2131	0.6679	0.5339	0.6418	0.7186	0.1731	0.9538	0.9985	0.8233
1	1.25	0.05	0.0501	0.0503	0.0506	0.0911	0.1385	0.1279	0.0946	0.1501	0.1438
		0.1	0.0541	0.0520	0.0541	0.1241	0.1983	0.1636	0.1377	0.2402	0.2162
		0.2	0.0528	0.0643	0.0741	0.1590	0.2492	0.1591	0.2034	0.3786	0.2939
		0.3	0.0584	0.0919	0.1053	0.1683	0.2413	0.1163	0.2501	0.4639	0.2970
		0.4	0.0671	0.1340	0.1263	0.1598	0.2038	0.0771	0.2759	0.5063	0.2438
	1.5	0.05	0.0502	0.0511	0.0523	0.2047	0.3648	0.3250	0.2200	0.4080	0.3849
		0.1	0.0517	0.0581	0.0659	0.3209	0.5361	0.4271	0.3751	0.6522	0.5878
		0.2	0.0612	0.1065	0.1432	0.4261	0.6404	0.3938	0.5717	0.8659	0.7357
		0.3	0.0832	0.2135	0.2547	0.4434	0.6076	0.2538	0.6778	0.9306	0.7255
		0.4	0.1169	0.3642	0.3147	0.4088	0.5063	0.1300	0.7237	0.9497	0.6052
	2	0.05	0.0508	0.0543	0.0590	0.5480	0.8204	0.7544	0.5961	0.8777	0.8489
		0.1	0.0565	0.0818	0.1121	0.7702	0.9474	0.8578	0.8573	0.9879	0.9708
		0.2	0.0943	0.2685	0.3856	0.7862	0.9744	0.7862	0.9739	0.9997	0.9933
		0.3	0.1801	0.5966	0.6523	0.5329	0.9555	0.5329	0.9918	1.0000	0.9892
		0.4	0.3052	0.8477	0.7228	0.2438	0.8862	0.2438	0.9951	1.0000	0.9497

Power calculations were performed using the genetic power calculator of Purcell & Sham (1). When the measured locus M occurs with frequency P_M then the ratio of frequencies between M and the disease locus D is given by $P_D : P_M$. The measure of linkage disequilibrium between M and P was set at $D' = 0.8$ or 1 . P_M was allowed to vary from 5% to 40%. P_D varied from 1.5-times lower to 1.5-

times higher than the marker allele frequency. A type I error rate of 0.05 (two-sided) was used for a sample size of $N=321$ and an unmatched case-control ratio of 1:1, assuming an overall disease prevalence rate of 0.0001.

Table II. Power calculations for main effects in the family-based study design

MAF	OR	Power		
		Recessive	Dominant	Multiplicative
0.05	1	0.05	0.05	0.05
	1.25	0.05076	0.1031	0.1108
	1.5	0.05172	0.2441	0.2744
	1.75	0.05287	0.4374	0.4957
	2	0.05426	0.6334	0.7072
	2.25	0.05587	0.791	0.8591
	2.5	0.05772	0.8954	0.9439
	2.75	0.05983	0.9538	0.9815
	3	0.06218	0.9818	0.9949
0.1	1	0.05	0.05	0.05
	1.25	0.05241	0.1398	0.1673
	1.5	0.05705	0.3665	0.4615
	1.75	0.06393	0.6277	0.759
	2	0.07304	0.8244	0.9265
	2.25	0.08437	0.932	0.9845
	2.5	0.09789	0.9779	0.9977
	2.75	0.1136	0.9938	0.9998
	3	0.1314	0.9985	1
0.15	1	0.05	0.05	0.05
	1.25	0.05567	0.1614	0.2162
	1.5	0.06895	0.4288	0.5945
	1.75	0.08975	0.7025	0.8788
	2	0.118	0.8781	0.9788
	2.25	0.1533	0.9591	0.9977
	2.5	0.1955	0.9882	0.9998
	2.75	0.2438	0.997	1
	3	0.2974	0.9993	1
0.2	1	0.05	0.05	0.05
	1.25	0.06109	0.1711	0.2573
	1.5	0.08955	0.4509	0.6842
	1.75	0.135	0.722	0.9326
	2	0.1967	0.8874	0.9924
	2.25	0.2724	0.9615	0.9995
	2.5	0.3588	0.9884	1
	2.75	0.4515	0.9968	1
	3	0.5455	0.9992	1
0.25	1	0.05	0.05	0.05
	1.25	0.06898	0.1716	0.2906
	1.5	0.1201	0.4462	0.7435
	1.75	0.202	0.7097	0.9578
	2	0.3098	0.874	0.9965
	2.25	0.4339	0.9521	0.9998
	2.5	0.562	0.9834	1
	2.75	0.6815	0.9945	1
	3	0.783	0.9983	1

0.3	1	0.05	0.05	0.05
	1.25	0.07946	0.1656	0.3163
	1.5	0.161	0.423	0.7818
	1.75	0.2891	0.6748	0.9702
	2	0.4472	0.8422	0.998
	2.25	0.6094	0.9306	0.9999
	2.5	0.751	0.9712	1
	2.75	0.8575	0.9884	1
	3	0.9269	0.9954	1
0.35	1	0.05	0.05	0.05
	1.25	0.09242	0.155	0.335
	1.5	0.2112	0.3869	0.8054
	1.75	0.3903	0.6219	0.9763
	2	0.5897	0.7915	0.9986
	2.25	0.7629	0.8926	1
	2.5	0.8832	0.9466	1
	2.75	0.9511	0.9739	1
	3	0.9827	0.9872	1
0.4	1	0.05	0.05	0.05
	1.25	0.1075	0.1417	0.347
	1.5	0.2685	0.3429	0.818
	1.75	0.4964	0.5547	0.9789
	2	0.7172	0.7211	0.9988
	2.25	0.8725	0.8329	1
	2.5	0.9541	0.9019	1
	2.75	0.9869	0.9427	1
	3	0.997	0.9663	1
0.45	1	0.05	0.05	0.05
	1.25	0.1242	0.1271	0.3527
	1.5	0.3297	0.2948	0.8219
	1.75	0.5972	0.4777	0.9792
	2	0.8168	0.6328	0.9987
	2.25	0.9379	0.7487	1
	2.5	0.9844	0.8295	1
	2.75	0.9971	0.8842	1
	3	0.9996	0.9207	1
0.5	1	0.05	0.05	0.05
	1.25	0.1418	0.1123	0.3524
	1.5	0.3908	0.2465	0.818
	1.75	0.6848	0.3965	0.9773
	2	0.886	0.5319	0.9985
	2.25	0.9716	0.642	0.9999
	2.5	0.9951	0.727	1
	2.75	0.9994	0.7909	1
	3	1	0.8385	1

Power calculations for the family-based study design were performed in PBAT (Version 3.5) (2). Recessive, dominant and multiplicative inheritance models were tested for minor allele frequencies (MAFs) varying from 0.05 to 0.5 and

main effect sizes (ORs) varying from 1.0 to 3.0. A type I error rate of 0.05 (two-sided) was used for a sample size consisting of 203 complete trios (father, mother and one affected child) and an overall disease prevalence rate of 0.0001.

Table III. Power calculations for gene-gene interaction effects in the case-control study design

Gene A	OR ^a	Gene B		
		Dominant (MAF=0.4;OR=1.0)	Recessive (MAF=0.3;OR=1.5)	Multiplicative (MAF=0.3;OR=3.0)
Dominant (MAF=0.3;OR=1.5)	1	0.05	0.05	0.05
	1.25	0.1008	0.072	0.1351
	1.5	0.2165	0.1253	0.3154
	1.75	0.3555	0.1973	0.5003
	2	0.489	0.2788	0.6452
	2.25	0.6025	0.3626	0.7458
	2.5	0.6927	0.4438	0.8125
	2.75	0.7619	0.5192	0.8563
	3	0.814	0.5871	0.8853
Recessive (MAF=0.2;OR=1.0)	1	0.05	0.05	0.05
	1.25	0.0584	0.0536	0.065
	1.5	0.0786	0.0629	0.1031
	1.75	0.1059	0.0762	0.1573
	2	0.1377	0.0924	0.2219
	2.25	0.172	0.111	0.2921
	2.5	0.2075	0.1312	0.3641
	2.75	0.2434	0.1528	0.4345
	3	0.2789	0.1753	0.5012
Multiplicative (MAF=0.1;OR=1.5)	1	0.05	0.05	0.05
	1.25	0.0945	0.0689	0.1283
	1.5	0.2007	0.1176	0.3155
	1.75	0.3358	0.1872	0.5286
	2	0.4727	0.2702	0.7027
	2.25	0.5944	0.3594	0.8201
	2.5	0.6941	0.4487	0.8913
	2.75	0.7717	0.5336	0.9325
	3	0.8302	0.6109	0.9559

Power calculations were performed for gene-gene interaction effects in Quanto (Version 1.1) (3). Different combinations of recessive, dominant and multiplicative (log-additive) inheritance models were tested for minor allele frequencies (MAFs) varying from 0.1 to 0.5 and main effect sizes (ORs) varying from 1.0 to 3.0. A type I error rate of 0.05 (two-sided) was used for a sample size of N=321 and an unmatched case-control ratio of 1:1, assuming an overall disease prevalence rate of 0.0001.

^a OR, interaction odds ratio.

POPULATION-BASED CASE-CONTROL ASSOCIATION STUDY

Statistical analyses were done using STATA/IC Version 10.1 (StataCorp, College Station, TX). Pearson's X^2 test or Fisher's exact test, as appropriate, was used to compare allele/genotype/haplotype carriership in patients and controls. Crude odds ratios (ORs) were measured using logistic regression or conventional 2x2 contingency table point estimation and are given with 95% confidence intervals (CIs). Haplotype reconstruction was performed using the FAMHAP Software, using parental data when available (4). Haplotype-specific ORs were estimated using the most common haplotype as reference and a likelihood ratio test implemented in FAMHAP was used to test for global haplotype association with disease status (5).

I assessed gender-specific associations through stratified analysis comparing male cases to male controls or female cases to female controls and the Mantel-Haenszel (MH) chi-square test of homogeneity was used to test for significant risk differences between males and females. I also tested for age-specific effects. The Shapiro-Francia test for normality was applied to the data and showed strong evidence against the assumption of normality for age at diagnosis. The nonparametric Kruskal-Wallis test was therefore used to compare the median age at diagnosis across genotype groups followed by Cuzick's nonparametric test for trend to assess gene-dosage effects. No significant effects of age at diagnosis were found for any of the variants tested therefore these results are not shown.

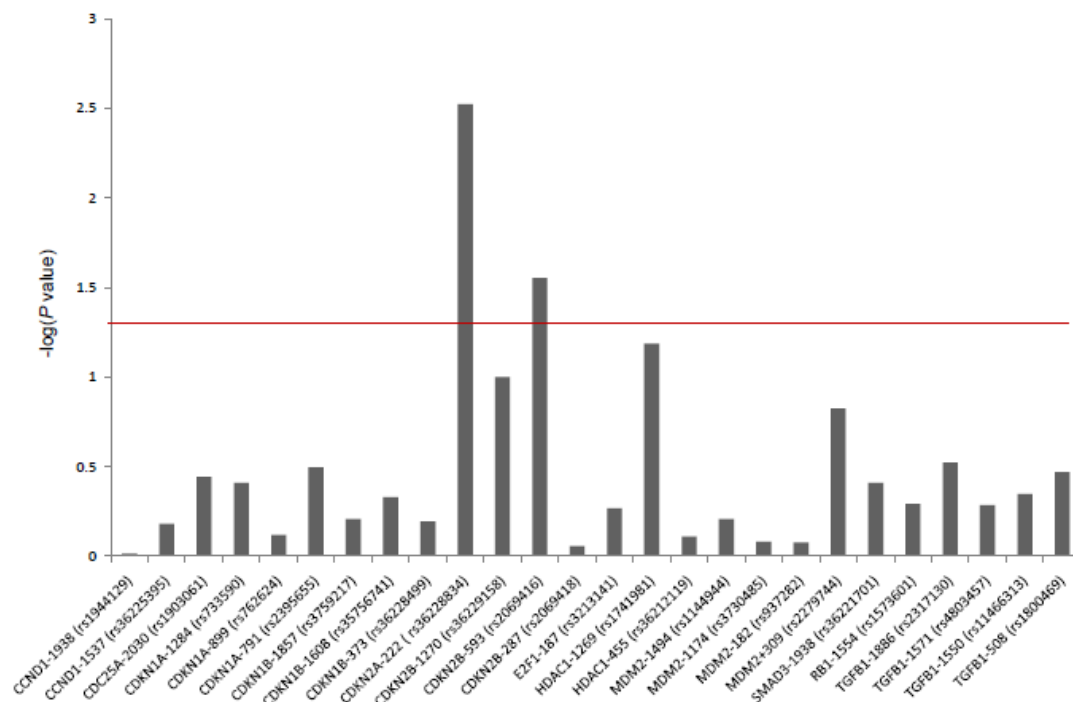


Figure I. Allelic associations of cell cycle checkpoint gene variants with childhood ALL

Results are presented as $-\log(P \text{ value})$ of the chi-square test for association for each variant. The horizontal reference line represents a nominal P value of 0.05. Note that for the tri-allelic variant *CDKN2B* -593A>T,C (rs2069416) only results for the T vs. A allele are shown.

Table IV. Allele frequencies of promoter SNPs in G1/S cell cycle checkpoint genes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk

Gene and SNP	Allele	No. (%)		OR (95% CI)	P
		ALL patients	Controls		
CCND1					
rs1944129	T	318 (51.1)	330 (51.2)	1.00 (.80-1.26)	.97
	C	304 (48.9)	314 (48.8)		
rs36225395*	-	333 (53.5)	356 (54.8)	1.05 (.84-1.32)	.66
	INS C	289 (46.5)	294 (45.2)		
CDC25A					
rs1903061	G	553 (91.3)	585 (89.7)	0.84 (.56-1.24)	.36
	T	53 (8.7)	67 (10.3)		
CDKN1A					
rs733590	T	370 (61.5)	406 (63.8)	1.11 (.87-1.40)	.39
	C	232 (38.5)	230 (36.2)		
rs762624	T	433 (73.6)	459 (72.9)	0.96 (.74-1.25)	.76
	G	155 (26.4)	171 (27.1)		
rs2395655	T	343 (57.6)	380 (60.3)	1.12 (.89-1.42)	.32
	C	253 (42.4)	250 (39.7)		
CDKN1B					
rs3759217	C	534 (89.3)	557 (88.4)	0.91 (.63-1.33)	.62
	T	64 (10.7)	73 (11.6)		
rs35756741	G	550 (90.2)	581 (91.4)	1.15 (.77-1.73)	.47
	A	60 (9.8)	55 (8.6)		
rs36228499	G	345 (57.9)	361 (56.6)	0.95 (.75-1.20)	.64
	T	251 (42.1)	277 (43.4)		
CDKN2A					
rs36228834	T	571 (93.0)	615 (96.7)	2.21 (1.26-3.96)	.003†
	A	43 (7.0)	21 (3.3)		
CDKN2B					
rs36229158	C	578 (95.4)	620 (97.2)	1.67 (.88-3.24)	.10†
	T	28 (4.6)	18 (2.8)		
rs2069416	A	393 (65.7)	376 (59.9)	1 (referent)	—
	T	188 (31.4)	235 (37.4)	0.77 (.60-.98)	.028
	C	17 (2.8)	17 (2.7)	0.96 (.45-2.03)	1.0†
rs2069418	G	345 (57.7)	362 (57.3)	0.98 (.78-1.24)	.88
	C	253 (42.3)	270 (42.7)		
E2F1					
rs3213141	C	463 (77.4)	480 (75.1)	0.92 (.70-1.21)	.54
	T	135 (22.6)	152 (24.1)		
HDAC1					
rs1741981	T	435 (71.8)	434 (67.0)	0.80 (.62-1.02)	.07
	C	171 (28.2)	214 (33.0)		

rs36212119	T	571 (91.2)	603 (91.6)	1.06 (.70-1.59)	.78
	C	55 (8.9)	55 (8.4)		
<i>MDM2</i>					
rs1144944	A	311 (49.4)	334 (50.8)	1.06 (.84-1.32)	.62
	G	319 (50.6)	324 (49.2)		
rs3730485	AAAAAGC(40bp) DEL 40bp	364 (58.9) 254 (41.1)	389 (59.5) 265 (40.5)	1.02 (.81-1.29)	.83
rs937282	C	320 (51.1)	339 (51.7)	1.02 (.82-1.28)	.84
	G	306 (48.9)	317 (48.3)		
rs2279744	T	421 (67.3)	412 (63.4)	0.84 (.66-1.07)	.15
	G	205 (32.7)	238 (36.6)		
<i>SMAD3</i>					
rs36221701	T	554 (88.8)	574 (87.2)	0.86 (.61-1.23)	.39
	C	70 (11.2)	84 (12.8)		
<i>RB1</i>					
rs1573601	C	462 (76.7)	478 (75.2)	0.92 (.70-1.20)	.51
	A	140 (23.3)	158 (24.8)		
<i>TGFB1</i>					
rs2317130	A	411 (65.9)	439 (68.6)	1.13 (.89-1.44)	.30
	G	213 (34.1)	201 (31.4)		
rs4803457	G	360 (58.8)	388 (60.6)	1.08 (.85-1.36)	.52
	A	252 (41.2)	252 (39.4)		
rs11466313	AGG DEL AGG	413 (67.0) 203 (33.0)	446 (69.0) 200 (31.0)	1.10 (.87-1.39)	.45
rs1800469	G	412 (66.2)	444 (68.7)	1.12 (.88-1.43)	.34
	A	210 (33.8)	202 (31.3)		

Percentages indicate number of individuals with a given allele/total number of genotyped alleles. The most common allele was used as reference. Association was tested for using a chi-square or Fisher's exact test, as appropriate.

Nominally significant results are shown in bold. OR indicates odds ratio; CI, confidence interval; —, not applicable; bp, base pairs.

*Variant *CCND1* -1537INS C (rs36225395) significantly deviated from Hardy-Weinberg equilibrium among controls ($P < 0.01$).

† P value indicates result of two-sided Fisher's exact test.

Table V. Distribution of G1/S cell cycle genotypes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk, as estimated by logistic regression

Gene and SNP	Genotype	No. (%)		OR (95% CI)	P
		ALL patients	Controls		
CCND1					
rs1944129	TT	78 (25.1)	74 (23.0)	1 (referent)	—
	TC	162 (52.1)	182 (56.5)	0.84 (.58-1.24)	.39
	CC	71 (22.8)	66 (20.5)	1.02 (.64-1.62)	.93
rs36225395*	-/-	85 (27.3)	85 (26.2)	1 (referent)	—
	-/C	163 (52.4)	186 (57.2)	0.88 (.61-1.26)	.48
	CC	63 (20.3)	54 (16.6)	1.17 (.73-1.87)	.52
CDC25A					
rs1903061	GG	253 (83.5)	262 (80.4)	1 (referent)	—
	GT	47 (15.5)	61 (18.7)	0.80 (.52-1.21)	.29
	TT	3 (1.0)	3 (0.9)	1.04 (.21-5.18)	.97
CDKN1A					
rs733590	TT	108 (35.9)	126 (39.6)	1 (referent)	—
	TC	154 (51.2)	154 (48.4)	1.17 (.83-1.64)	.89
	CC	39 (12.9)	38 (11.9)	1.20 (.72-2.00)	.68
rs762624	TT	160 (54.4)	168 (53.3)	1 (referent)	—
	TG	113 (38.4)	123 (39.0)	0.96 (.69-1.35)	.83
	GG	21 (7.1)	24 (7.6)	0.92 (.49-1.71)	.79
rs2395655‡	TT	94 (31.5)	112 (35.6)	1 (referent)	—
	TC	155 (52.0)	156 (49.5)	1.18 (.83-1.68)	.35
	CC	49 (16.5)	47 (14.9)	1.24 (.76-2.02)	.38
CDKN1B					
rs3759217	CC	238 (79.6)	243 (77.1)	1 (referent)	—
	CT	58 (19.4)	71 (22.5)	0.83 (.56-1.23)	.36
	TT	3 (1.0)	1 (0.3)	3.06 (.32-29.65)	.37†
rs35756741	GG	245 (80.3)	267 (84.0)	1 (referent)	—
	GA	60 (19.7)	47 (14.8)	1.39 (.91-2.11)	.12
	AA	0 (0.0)	4 (1.2)	—	—
rs36228499	GG	93 (31.2)	99 (31.0)	1 (referent)	—
	GT	159 (53.4)	163 (51.1)	1.04 (.73-1.48)	.84
	TT	46 (15.4)	57 (17.9)	0.86 (.53-1.39)	.54
CDKN2A					
rs36228834	TT	266 (86.6)	298 (93.7)	1 (referent)	—
	TA	39 (12.7)	19 (6.0)	2.30 (1.30-4.08)	.004
	AA	2 (0.7)	1 (0.3)	2.24 (.20-24.85)	.60†
CDKN2B					
rs36229158	CC	277 (91.4)	302 (94.7)	1 (referent)	—
	CT	24 (7.9)	16 (5.0)	1.64 (.85-3.14)	.14
	TT	2 (0.7)	1 (0.3)	2.18 (.20-24.18)	.61†

rs2069416	AA	129 (43.1)	107 (34.1)	1 (referent)	—
	AT	124 (41.5)	150 (47.8)	0.69 (.48-.97)	.03
	AC	11 (3.7)	12 (3.8)	0.60 (.35-1.03)	.07†
	TT	29 (9.7)	40 (12.7)	0.76 (.32-1.79)	.53
	TC	6 (2.0)	5 (1.6)	0.99 (.30-3.35)	.99†
	CC	0	0	—	—
rs2069418	GG	102 (34.1)	97 (30.7)	1 (referent)	—
	GC	141 (47.2)	168 (53.2)	0.80 (.56-1.14)	.22
	CC	56 (18.7)	51 (16.1)	1.04 (.65-1.67)	.86
<i>E2F1</i>					
rs3213141	CC	172 (57.5)	180 (57.0)	1 (referent)	—
	CT	119 (29.8)	120 (38.0)	1.04 (.75-1.44)	.82
	TT	8 (2.7)	16 (5.1)	0.52 (.22-1.25)	.15†
<i>HDAC1</i>					
rs1741981	TT	151 (49.8)	143 (44.1)	1 (referent)	—
	TC	133 (43.9)	148 (45.7)	0.85 (.61-1.18)	.33
	CC	19 (6.3)	33 (10.2)	0.54 (.30-1.00)	.05
rs36212119	TT	261 (83.4)	277 (84.2)	1 (referent)	—
	TC	49 (15.6)	49 (14.9)	1.06 (.69-1.63)	.79
	CC	3 (1.0)	3 (0.9)	1.06 (.21-5.30)	.94†
<i>MDM2</i>					
rs1144944	AA	73 (23.2)	82 (24.9)	1 (referent)	—
	AG	165 (52.4)	170 (51.7)	1.09 (.74-1.60)	.66
	GG	77 (24.4)	77 (23.4)	1.12 (.72-1.76)	.61
rs3730485	AAAAAGC(40bp)	104 (33.6)	119 (36.4)	1 (referent)	—
	AAAAAGC(40bp)/-	156 (50.5)	151 (46.2)	1.18 (.84-1.67)	.34
	-/-	49 (15.9)	57 (17.4)	0.98 (.62-1.56)	.94
rs937282	CC	78 (24.9)	84 (25.6)	1 (referent)	—
	CG	164 (52.4)	171 (52.1)	1.03 (.71-1.50)	.87
	GG	71 (22.7)	73 (22.3)	1.05 (.67-1.64)	.84
rs2279744	TT	141 (45.1)	125 (38.5)	1 (referent)	—
	TG	139 (44.4)	162 (49.8)	0.76 (.55-1.06)	.11
	GG	33 (10.5)	38 (11.7)	0.77 (.45-1.30)	.33
<i>SMAD3</i>					
rs36221701	TT	246 (78.8)	254 (77.2)	1 (referent)	—
	TC	62 (19.9)	66 (20.1)	0.97 (.66-1.43)	.88
	CC	4 (1.3)	9 (2.7)	0.46 (.14-1.51)	.20†
<i>RB1</i>					
rs1573601	CC	175 (58.1)	176 (55.3)	1 (referent)	—
	CA	112 (37.2)	126 (39.6)	0.89 (.64-1.24)	.51
	AA	14 (4.7)	16 (5.0)	0.88 (.42-1.86)	.74†
<i>TGFB1</i>					
rs2317130	AA	138 (44.2)	144 (50.0)	1 (referent)	—
	AG	135 (43.3)	151 (47.2)	0.93 (.67-1.30)	.68
	GG	39 (12.5)	25 (7.8)	1.63 (.94-2.83)	.09
rs4803457	GG	105 (34.3)	113 (35.3)	1 (referent)	—
	GA	150 (49.0)	162 (50.6)	1.00 (.70-1.41)	.98
	AA	51 (16.7)	45 (14.1)	1.22 (.75-1.97)	.42
rs11466313	AGG/AGG	140 (45.4)	149 (46.1)	1 (referent)	—

	AGG/-	133 (43.2)	148 (45.8)	0.96 (.69-1.33)	.79
	-/-	35 (11.4)	26 (8.1)	1.43 (.82-2.50)	.21
rs1800469	GG	138 (44.4)	147 (45.5)	1 (referent)	—
	GA	136 (43.7)	150 (46.4)	.97 (.70-1.34)	.84
	AA	37 (11.9)	26 (8.1)	1.52 (.87-2.63)	.14

Percentages indicate number of individuals with a given genotype/total number of genotyped individuals. The homozygous genotype of the most common allele was used as reference. Association was tested for using a chi-square or Fisher's exact test, as appropriate. Nominally significant results are shown in bold. OR indicates odds ratio; CI, confidence interval; —, not applicable; bp, base pairs.

*Variant *CCND1* -1537INS C (rs36225395) significantly deviated from Hardy-Weinberg equilibrium among controls ($P < 0.01$).

† P value indicates result of two-sided Fisher's exact test.

‡ Significant risk differences between males and females for variant *CDKN1A* rs2395655 CC vs. TT; Males: OR(95%CI)= 0.77(.41-1.45), Females: OR(95%CI)= 2.36(1.09-5.08), Mantel-Haenszel $P = 0.03$.

Table VI. Distribution of G1/S cell cycle haplotypes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk

Gene and Haplotype	DNA variant			No. (%)		OR (95% CI)	P	Global χ^2 (df)	Global P
				ALL patients	Controls				
<i>CCND1</i>	rs1944129	rs36225395							
CCND1*1	T	DEL	C	319 (51.1)	337 (51.5)	1 (referent)	—	1.57 (3)	.67
CCND1*2	C		C	287 (46.0)	297 (45.4)	1.02 (.81-1.28)	.86		
CCND1*3	C	DEL	C	16 (2.6)	20 (3.1)	0.85 (.40-1.75)	.73†		
CCND1*4	T		C	2 (0.3)	0 (0.0)	—	—		
<i>CDKN1A</i>	rs733590	rs762624	rs2395655						
CDKN1A*1	T	T	T	346 (56.1)	370 (58.2)	1 (referent)	—	7.54 (7)	.37
CDKN1A*2	C	G	C	126 (20.5)	140 (22.0)	0.96 (.72-1.29)	.79		
CDKN1A*3	C	T	C	106 (17.2)	86 (13.5)	1.32 (.95-1.84)	.09		
CDKN1A*4	T	G	C	29 (4.7)	29 (4.6)	1.07 (.60-1.89)	.89†		
CDKN1A*	—	—	—	9 (1.5)	11 (1.7)	0.87 (.32-2.35)	.82†		
<i>CDKN1B</i>	rs3759217	rs35756741	rs36228499						
CDKN1B*1	C	G	T	266 (42.8)	277 (43.4)	1 (referent)	—	4.01 (5)	.55
CDKN1B*2	C	G	G	231 (37.1)	233 (36.5)	1.03 (.80-1.33)	.80		
CDKN1B*3	T	G	G	62 (10.0)	73 (11.4)	0.88 (.59-1.31)	.52		
CDKN1B*4	C	A	G	60 (9.6)	55 (8.6)	1.14 (.74-1.74)	.53		
CDKN1B*5	—	—	—	3 (0.5)	0 (0.0)	—	—		
<i>CDKN2B</i>	rs36229158	rs2069416	rs2069418						
CDKN2B*1	C	A	C	267 (43.1)	276 (43.3)	1 (referent)	—	20.88 (7)	.004
CDKN2B*2	C	T	G	192 (31.0)	236 (37.0)	0.84 (.65-1.09)	.19		
CDKN2B*3	C	A	G	116 (18.7)	91 (14.3)	1.32 (.94-1.84)	.09		
CDKN2B*4	T	A	G	29 (4.7)	18 (2.8)	1.67 (.87-3.26)	.13†		
CDKN2B*	—	—	—	16 (2.6)	17 (2.7)	0.97 (.45-2.09)	1.0†		
<i>HDAC1</i>	rs1741981	rs36212119							
HDAC1*1	T	T		406 (63.8)	390 (59.3)	1 (referent)	—	5.25 (3)	0.15
HDAC1*2	C	T		172 (27.0)	213 (32.4)	0.76 (.60-1.00)	.04		
HDAC1*3	T	C		58 (9.2)	54 (8.2)	1.03 (.68-1.56)	.88		
HDAC1*4	C	C		0 (0.0)	1 (0.1)	—	—		

<i>MDM2</i>	rs1144944	rs3730485	rs937282	rs2279744						
MDM2*1	G	DEL 40bp	G	T	254 (39.8)	261 (39.7)	1 (referent)	—		
MDM2*2	A	AAAAAGC(40bp)	C	G	196 (30.7)	238 (36.2)	0.85 (.65-1.10)	.20		
MDM2*3	A	AAAAAGC(40bp)	C	T	112 (17.6)	92 (14.0)	1.25 (.89-1.76)	.18	16.68	.27
MDM2*4	G	AAAAAGC(40bp)	G	T	52 (8.5)	54 (8.2)	0.99 (.64-1.54)	.96	(14)	
MDM2*	—	—	—	—	24 (3.8)	13 (2.0)	1.90 (.90-4.15)	.09†		
<i>TGFB1</i>	rs2317130	rs4803457	rs11466313	rs1800469						
TGFB1*1	A	G	AGG	G	362 (56.9)	387 (59.9)	1 (referent)	—		
TGFB1*2	G	A	DEL AGG	A	205 (32.2)	196 (30.3)	1.12 (.87-1.44)	.37	10.77	.55
TGFB1*3	A	A	AGG	G	49 (7.7)	52 (8.1)	1.01 (.65-1.56)	.97	(12)	
TGFB1*	—	—	—	—	20 (3.1)	11 (1.7)	1.94 (.87-4.55)	.10†		

Haplotype reconstruction was performed using the FAMHAP Software, using parental data when available. Percentages indicate number of chromosomes with given haplotype/total number of chromosomes for each gene. Haplotypes with relative frequencies <5% are grouped and are represented as * combinations of the respective DNA variants. The risk of ALL was evaluated for each haplotype compared with the most common haplotype which was chosen as reference.

Association was tested for using a chi-square or Fisher's exact test, as appropriate. A likelihood ratio test was performed in FAMHAP to compare global haplotype differences between cases and controls and is reported here as a Global chi-square (X^2) test with number of haplotype parameters different from zero—1 degrees of freedom. Nominally significant results are shown in bold. OR indicates odds ratio; CI, confidence interval; —, not applicable; bp, base pairs.

†*P* value indicates result of two-sided Fisher's exact test.

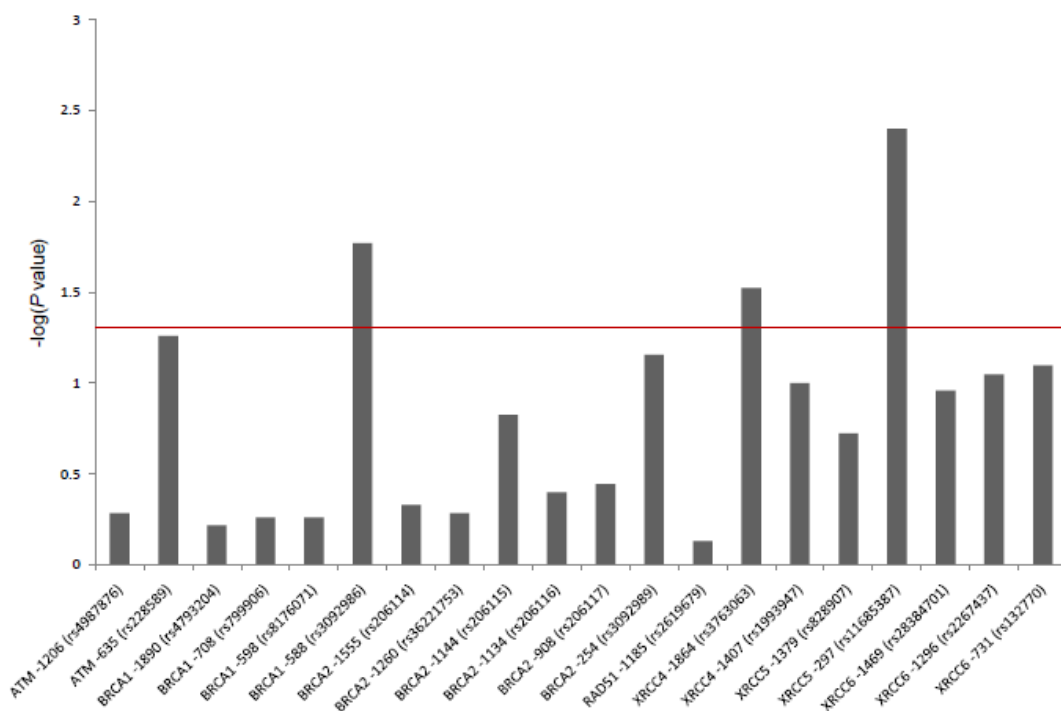


Figure II. Allelic associations of DNA double-strand break repair gene variants with childhood ALL

Results are presented as $-\log(P \text{ value})$ of the chi-square test for association for each variant. The horizontal reference line represents a nominal P value of 0.05.

Table VII. Allele frequencies of promoter SNPs in DNA double-strand break repair genes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk

Gene and SNP	Allele	No. (%)		OR (95% CI)	P
		ALL patients	Controls		
ATM					
rs4987876	G	516 (89.3)	573 (90.4)	1.13 (.76-1.67)	.52
	T	62 (10.7)	61 (9.6)		
rs228589	T	359 (63.7)	368 (58.2)	0.80 (.63-1.01)	.06
	A	205 (36.3)	264 (41.8)		
BRCA1					
rs4793204	T	404 (66.2)	434 (67.6)	1.06 (.83-1.36)	.61
	C	206 (33.8)	208 (32.4)		
rs799906	A	394 (65.0)	433 (66.6)	1.07 (.84-1.35)	.55
	G	212 (35.0)	217 (33.4)		
rs8176071	-	405 (66.2)	443 (67.7)	1.07 (.85-1.37)	.55
	INS ACA	207 (33.8)	211 (32.3)		
rs3092986	A	575 (94.6)	590 (91.0)	0.58 (.36-.99)	.017†
	G	33 (5.4)	58 (9.0)		
BRCA2					
rs206114	C	357 (56.7)	386 (58.7)	1.09 (.86-1.36)	.47
	A	273 (43.3)	272 (41.3)		
rs3072036	GTCTAA	367 (58.6)	395 (60.4)	1.08 (.86-1.35)	.52
	DEL GTCTAA	259 (41.4)	259 (39.6)		
rs206115	A	328 (55.0)	377 (59.1)	1.18 (.94-1.49)	.15
	G	268 (45.0)	261 (40.9)		
rs206116	C	332 (57.4)	378 (59.8)	1.10 (.87-1.40)	.40
	T	246 (42.6)	254 (40.2)		
rs206117	C	351 (57.7)	387 (60.3)	1.11 (.88-1.40)	.36
	T	257 (42.3)	255 (39.7)		
rs3092989	G	529 (83.4)	523 (79.5)	0.77 (.57-1.03)	.07
	A	105 (16.6)	135 (20.5)		
RAD51					
rs2619679	A	313 (50.0)	330 (50.9)	1.04 (.83-1.30)	.74
	T	313 (50.0)	318 (49.1)		
XRCC4					
rs3763063	T	336 (56.6)	324 (50.5)	0.78 (.62-.98)	.03
	C	258 (43.4)	318 (49.5)		
rs1993947	C	517 (88.8)	581 (91.6)	1.38 (.94-2.06)	.10
	G	65 (11.2)	53 (8.4)		

XRCC5						
rs828907	G	355 (56.5)	347 (52.9)	.86 (.69-1.08)		.19
	T	273 (43.5)	309 (47.1)			
rs11685387	C	419 (67.1)	488 (74.4)	1.42 (1.11-1.82)		.004
	T	205 (32.9)	168 (25.6)			
XRCC6						
rs28384701*	C	575 (93.3)	615 (95.5)	1.51 (.90-2.56)		.11†
	T	41 (6.7)	29 (4.5)			
rs2267437	C	349 (56.6)	393 (61.4)	1.22 (.97-1.52)		.09
	G	267 (43.4)	247 (38.6)			
rs132770	G	498 (79.0)	490 (74.9)	0.79 (.60-1.04)		.08
	A	132 (21.0)	164 (25.1)			

Percentages indicate number of individuals with a given allele/total number of genotyped individuals. The most common allele was used as reference.

Association was tested for using a chi-square or Fisher's exact test, as appropriate. Nominally significant results are shown in bold. OR indicates odds ratio; CI, confidence interval; —, not applicable.

*Variant *XRCC6* -1469C>T (rs28384701) significantly deviated from Hardy-Weinberg equilibrium among controls ($P<0.01$).

† P value indicates result of two-sided Fisher's exact test.

Table VIII. Distribution of DNA double-strand break repair genotypes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk, as estimated by logistic regression

Gene and SNP	Genotype	No. (%)		OR (95% CI)	P
		ALL patients	Controls		
ATM					
rs4987876	GG	231 (79.9)	261 (82.3)	1 (referent)	—
	GT	54 (18.7)	51 (16.1)	1.20 (.78-1.82)	.29
	TT	4 (1.4)	5 (1.6)	0.90 (.24-3.41)	1†
rs228589	TT	111 (39.4)	111 (35.1)	1 (referent)	—
	TA	137 (48.6)	146 (46.2)	0.94 (.66-1.33)	.68
	AA	34 (12.0)	59 (18.7)	0.58 (.35-.95)	.03
BRCA1					
rs4793204	TT	130 (42.6)	149 (46.4)	1 (referent)	—
	TC	144 (47.2)	136 (42.4)	1.21 (.87-1.69)	.25
	CC	31 (10.2)	36 (11.2)	0.99 (.58-1.68)	.96
rs799906	AA	124 (40.9)	145 (44.6)	1 (referent)	—
	AG	146 (48.2)	143 (44.0)	1.19 (.86-1.66)	.30
	GG	33 (10.9)	37 (11.4)	1.04 (.61-1.77)	.88
rs8176071	-/-	130 (42.5)	151 (46.2)	1 (referent)	—
	-/ACA	145 (47.4)	141 (43.1)	1.19 (.86-1.66)	.29
	ACA/ACA	31 (10.1)	35 (10.7)	1.03 (.60-1.76)	.92
rs3092986	AA	272 (89.5)	268 (82.7)	1 (referent)	—
	AG	31 (10.2)	54 (16.7)	0.56 (.35-.91)	.018
	GG	1 (0.3)	2 (0.6)	0.49 (.04-5.47)	.62†
BRCA2					
rs206114	CC	112 (35.5)	116 (35.3)	1 (referent)	—
	CA	133 (42.2)	154 (46.8)	0.89 (.63-1.27)	.53
	AA	70 (22.2)	59 (17.9)	1.23 (.80-1.89)	.35
rs36221753	GTCTAA/GTCTAA	115 (36.7)	117 (35.8)	1 (referent)	—
	GTCTAA/-	137 (43.8)	161 (49.2)	.86 (.61-1.22)	.41
	-/-	61 (19.5)	49 (15.0)	1.27 (.80-2.00)	.31
rs206115	AA	97 (32.5)	110 (34.5)	1 (referent)	—
	AG	134 (45.0)	157 (49.2)	.97 (.68-1.38)	.86
	GG	67 (22.5)	52 (16.3)	1.46 (.93-2.30)	.10
rs206116	CC	105 (36.3)	111 (35.1)	1 (referent)	—
	CT	122 (42.2)	156 (49.4)	0.83 (.58-1.18)	.30
	TT	62 (21.5)	49 (15.5)	1.34 (.84-2.12)	.21
rs206117	CC	108 (35.5)	115 (35.8)	1 (referent)	—
	CT	135 (44.4)	157 (48.9)	0.91 (.64-1.30)	.62
	TT	61 (20.1)	49 (15.3)	1.32 (.84-2.10)	.23

rs3092989	GG	220 (69.4)	205 (62.3)	1 (referent)	—
	GA	89 (28.1)	113 (34.3)	0.73 (.52-1.03)	.07
	AA	8 (2.5)	11 (3.3)	0.68 (.27-1.72)	.49†
<i>RAD51</i>					
rs2619679	AA	76 (24.3)	81 (25.0)	1 (referent)	—
	AT	161 (51.4)	168 (51.9)	1.02 (.70-1.49)	.91
	TT	76 (24.3)	75 (23.1)	1.08 (.69-1.69)	.74
<i>XRCC4</i>					
rs3763063	TT	98 (33.0)	79 (24.6)	1 (referent)	—
	TC	140 (47.1)	166 (51.7)	0.68 (.47-.98)	.042
	CC	59 (19.9)	76 (23.7)	0.63 (.40-.98)	.042
rs1993947	CC	229 (78.7)	266 (83.9)	1 (referent)	—
	CG	59 (20.3)	49 (15.5)	1.40 (.92-2.12)	.12
	GG	3 (1.0)	2 (0.6)	1.74 (.29-10.52)	.67†
<i>XRCC5</i>					
rs828907‡	GG	101 (32.2)	88 (26.8)	1 (referent)	—
	GT	153 (48.7)	171 (52.1)	0.78 (.54-1.12)	.17
	TT	60 (19.1)	69 (21.1)	0.76 (.48-1.19)	.22
rs11685387	CC	140 (44.9)	178 (54.3)	1 (referent)	—
	CT	139 (44.5)	132 (40.2)	1.34 (.97-1.85)	.08
	TT	33 (10.6)	18 (5.5)	2.33 (1.26-4.31)	.007†
<i>XRCC6</i>					
rs28384701*	CC	267 (86.7)	297 (92.2)	1 (referent)	—
	CT	41 (13.3)	21 (6.5)	2.17 (1.25-3.77)	.007†
	TT	0 (0.0)	4 (1.2)	—	—
rs2267437	CC	97 (31.5)	115 (35.9)	1 (referent)	—
	CG	155 (50.3)	163 (50.9)	1.13 (.80-1.60)	.50
	GG	56 (18.2)	42 (13.1)	1.58 (.98-2.56)	.06
rs132770	GG	197 (62.5)	189 (57.8)	1 (referent)	—
	GA	104 (33.0)	112 (34.2)	0.89 (.64-1.24)	.50
	AA	14 (4.5)	26 (8.0)	0.52 (.26-1.02)	.07†

Percentages indicate number of individuals with a given genotype/total number of genotyped individuals. The homozygous genotype of the most common allele was used as reference. Association was tested for using a chi-square or Fisher's exact test, as appropriate. Nominally significant results are shown in bold.

OR indicates odds ratio; CI, confidence interval; —, not applicable.

*Variant *XRCC6* -1469C>T (rs28384701) significantly deviated from Hardy-Weinberg equilibrium among controls ($P<0.01$).

† P value indicates result of two-sided Fisher's exact test.

‡ Significant risk differences between males and females for variant *XRCC5* rs828907 TT vs. GG; Males: OR(95%CI)= 0.44(.24-.81), Females: OR(95%CI)= 1.56(.79-3.10), Mantel-Haenszel $P= 0.007$.

Table IX. Distribution of DNA double-strand break repair haplotypes among B-cell ALL cases and controls from the Quebec childhood ALL cohort and their effect on ALL risk

Gene and Haplotype		DNA variant					No. (%)		OR (95% CI)	P	Global χ^2 (df)	Global P
							ALL patients	Controls				
ATM	rs4987876	rs228589										
ATM*1	G	T					348 (55.8)	321 (50.0)	1 (referent)	—		
ATM*2	G	A					212 (34.0)	260 (40.5)	0.75 (.59-.96)	.018	2.41 (3)	.49
ATM*3	T	T					62 (9.9)	57 (8.9)	1.00 (.67-1.51)	.99		
ATM*4	T	A					2 (0.3)	4 (0.6)	0.46 (.04-3.25)	.44†		
BRCA1	rs4793204	rs799906	rs8176071	rs3092986								
BRCA1*1	T	A	DEL ACA	A			377 (60.4)	377 (57.6)	1 (referent)	—		
BRCA1*2	C	G	ACA	A			209 (33.5)	211 (32.3)	0.99 (.77-1.27)	.94	9.77 (7)	.20
BRCA1*3	T	A	DEL ACA	G			32 (5.1)	58 (8.9)	0.55 (.34-.89)	.01†		
BRCA1*	—	—	—	—			6 (1.0)	8 (1.2)	0.75 (.21-2.49)	.79†		
BRCA2	rs206114	rs36221753	rs206115	rs206116	rs206117	rs3092989						
BRCA2*1	A	GTCTAA	G	C	C	G	261 (41.2)	263 (40.0)	1 (referent)	—		
BRCA2*2	C	DEL GTCTAA	A	T	T	G	246 (38.8)	245 (37.2)	1.01 (.78-1.30)	.93	41.23 (22)	.008
BRCA2*3	C	GTCTAA	A	C	C	A	89 (14.0)	125 (19.0)	0.72 (.51-1.00)	.04		
BRCA2*	—	—	—	—	—	—	38 (6.0)	25 (3.8)	1.53 (.87-2.73)	.14†		
XRCC4	rs6452505	rs1993947										
XRCC4*1	C	C					261 (41.5)	320 (49.7)	1 (referent)	—	5.58 (3)	.13
XRCC4*2	T	C					293 (46.7)	271 (42.1)	1.33 (1.04-1.68)	.02		
XRCC4*3	T	G					73 (11.6)	53 (8.2)	1.69 (1.12-2.55)	.008		
XRCC4*4	—	—					1 (0.2)	0 (0.0)	—	—		

XRCC5	rs828907	rs11685387							
XRCC5*1	T	C		280 (43.9)	309 (47.1)	1 (referent)	—	9.36 (2)	.009
XRCC5*2	G	C		146 (22.9)	179 (27.3)	0.90 (.68-1.19)	.45		
XRCC5*3	G	T		212 (33.2)	168 (25.6)	1.39 (1.06-1.82)	.01		
XRCC6	rs28384701	rs2267437	rs132770						
XRCC6*1	C	G	G	266 (41.8)	253 (38.7)	1 (referent)	—	17.07 (8)	.03
XRCC6*2	C	C	G	224 (35.2)	231 (35.3)	0.92 (.71-1.20)	.53		
XRCC6*3	C	C	A	90 (14.2)	137 (20.9)	0.62 (.45-.87)	.003		
XRCC6*4	T	C	A	42 (6.6)	27 (4.1)	1.48 (.86-2.57)	.16†		
XRCC6*	—	—	—	14 (2.2)	6 (0.9)	2.22 (.78-7.15)	.11†		

Haplotype reconstruction was performed using the FAMHAP Software, using parental data when available. Percentages indicate number of chromosomes with given haplotype/total number of chromosomes for each gene. Haplotypes with relative frequencies <5% are grouped and are represented as * combinations of the respective DNA variants. The risk of ALL was evaluated for each haplotype compared with the most common haplotype which was chosen as reference. Association was tested for using a chi-square or Fisher's exact test, as appropriate. A likelihood ratio test was performed in FAMHAP to compare global haplotype differences between cases and controls and is reported here as a Global chi-square (X^2) test with number of haplotype parameters different from zero—1 degrees of freedom. Nominally significant results are shown in bold. OR indicates odds ratio; CI, confidence interval; —, not applicable; bp, base pairs.

†P value indicates result of two-sided Fisher's exact test.

FAMILY-BASED ASSOCIATION STUDY

Transmission disequilibrium from parents to children of individual SNPs and corresponding haplotypes was assessed with the FBAT (family-based association test) software (6, 7). A multiallelic test was also carried out in FBAT to obtain the global haplotype association significance in the family-based setting.

Please note that multiple testing corrections were not performed on the data presented here and only nominal P values are shown.

Table X. Family-based association analysis of promoter SNPs in G1/S cell cycle checkpoint genes

Gene and SNP	DNA variant	MAF	No. of families*	S-E(S)†	Var(S)‡	Z	P
<i>CCND1</i>							
rs1944129	T>C	.49	123	-6.00	40.00	-0.95	.34
rs36225395	INS C	.47	120	-2.50	37.75	0.41	.68
<i>CDC25A</i>							
rs1903061	G>T	.08	41	1.50	10.75	0.46	.65
<i>CDKN1A</i>							
rs733590	T>C	.39	114	1.50	38.75	0.24	.81
rs762624	T>G	.28	101	0.00	32.50	0.00	1
rs2395655	T>C	.44	111	5.00	38.00	0.81	.42
<i>CDKN1B</i>							
rs3759217	C>T	.12	67	-5.50	19.25	-1.25	.21
rs35756741	G>A	.10	58	3.00	15.00	0.78	.44
rs36228499	G>T	.42	102	-5.00	35.50	-0.84	.40
<i>CDKN2A</i>							
s36228834	T>A	.06	33	11.00	8.50	3.77	.0002
<i>CDKN2B</i>							
rs36229158	C>T	.04	23	7.00	6.00	2.86	.004
rs2069416	A>T,C	.35/.03	117/15	17.50	37.75/4.00	-2.85/-0.5	.004/.62
rs2069418	G>C	.42	124	5.00	41.50	0.78	.44
<i>E2F1</i>							
rs3213141	C>T	.24	100	-0.50	30.25	-0.09	.93
<i>HDAC1</i>							
rs1741981	T>C	.29	100	-7.50	32.25	-1.32	.19
rs36212119	T>C	.09	30	2.50	7.75	0.98	.37
<i>MDM2</i>							
rs1144944	A>G	.50	125	1.50	43.75	0.23	.82
rs3730485	DEL AAAAAAGC(40bp)	.41	125	-9.50	41.25	-1.48	.14
rs937282	C>G	.48	121	-5.50	42.25	-0.85	.40
rs2279744	T>G	.34	126	-2.00	41.00	-0.31	.75
<i>SMAD3</i>							
rs36221701	T>C	.11	47	-2.50	13.75	-0.67	.50
<i>RB1</i>							
rs1573601	C>A	.23	94	-2.00	29.00	-0.37	.71
<i>TGFB1</i>							
rs2317130	A>G	.33	113	5.00	36.50	0.83	.41
rs4803457	G>A	.40	118	9.00	39.00	1.44	.15
rs11466313	DEL AGG	.32	112	6.50	35.25	1.10	.27
rs1800469	G>A	.33	119	2.00	38.00	0.32	.75

FBAT analyses were performed under the additive model. Only the results for the minor alleles are shown. A positive Z value indicates overtransmission of a given allele to affected offspring; a negative Z value indicates

undertransmission. Nominally significant results are shown in bold. MAF indicates minor (i.e., variant) allele frequency.

*Number of informative families (i.e., families with at least one heterozygote parent).

†Test statistic (S) from FBAT for the observed number of transmitted alleles – the expected value of S ($E(S)$) under Mendelian transmission and the null hypothesis (i.e., no linkage or association).

‡Variance of the test statistic S .

Table XI. Family-based association analysis of promoter haplotypes in G1/S cell cycle checkpoint genes

Gene and Haplotype	Alleles	Freq.	No. of families*	S-E(S)†	Var(S)‡	Z	P	Global χ^2 (df)	Global P
CCND1									
CCND1*1	T(-)	.49	102	5.00	39.00	0.80	.42	2.38 (3)	.50
CCND1*2	CC	.47	102	-2.50	36.75	-0.41	.68		
CCND1*3	C(-)	.03	14	-2.00	3.50	-1.07	.29		
CCND1*4	TC	.001	0	—	—	—	—		
CDKN1A									
CDKN1A*1	TTT	.59	88	-7.00	37.38	-1.15	.25	5.09 (4)	.28
CDKN1A*2	CGC	.20	78	-4.90	27.81	-0.94	.35		
CDKN1A*3	CTC	.15	71	7.90	21.38	1.72	.09		
CDKN1A*4	TGC	.04	31	3.40	9.18	1.13	.26		
CDKN1B									
CDKN1B*1	CGT	.43	93	-3.47	33.76	-0.60	.55	3.90 (4)	.42
CDKN1B*2	CGG	.39	102	6.97	32.21	1.23	.22		
CDKN1B*3	TGG	.10	55	-5.97	15.71	-1.51	.13		
CDKN1B*4	CAG	.08	49	2.50	12.25	0.71	.48		
CDKN2B									
CDKN2B*1	CAC	.42	109	5.91	40.49	0.93	.35	14.32 (6)	.03
CDKN2B*2	CTG	.35	106	-16.51	41.21	-2.57	.01		
CDKN2B*3	CAG	.18	77	3.59	21.13	0.78	.43		
CDKN2B*4	TAG	.03	22	6.42	6.18	2.59	.01		
HDAC1									
HDAC1*1	TT	.65	90	4.34	27.68	0.82	.41	3.93 (3)	.27
HDAC1*2	CT	.28	81	-5.84	25.69	-1.15	.25		
HDAC1*3	TC	.07	26	2.66	5.77	1.11	.27		
HDAC1*4	CC	.002	0	—	—	—	—		
MDM2									
MDM2*1	G(-)GT	.40	93	-5.50	34.68	-0.94	.35	8.15 (4)	.09
MDM2*2	A(40bp)CG	.32	96	1.09	35.91	0.18	.86		
MDM2*3	A(40bp)CT	.16	77	2.92	20.97	0.64	.52		
MDM2*4	G(40bp)GT	.07	39	5.58	9.73	1.79	.07		
TGFB1									
TGFB1*1	AG(AGG)G	.63	107	-6.01	41.45	-0.93	.35	5.93 (3)	.12
TGFB1*2	GA(-)A	.27	93	6.99	33.42	1.21	.23		
TGFB1*3	AA(AGG)G	.07	39	2.50	9.63	0.81	.42		

Haplotype-specific FBAT analyses were performed under the additive model. A multiallelic chi-square test (χ^2) was also carried out in FBAT to obtain the global haplotype association significance in the family-based setting. A positive Z value indicates overtransmission of a given haplotype to affected offspring; a negative Z value indicates undertransmission. Nominally significant results are shown in

bold. Freq. indicates haplotype frequency; df indicates degrees of freedom; —, not applicable due to lack of informative families (i.e., < 10).

*Number of informative families (i.e., families with at least 1 heterozygote parent).

†Test statistic (S) from FBAT for the observed number of transmitted haplotypes – expected value of S (E(S)) under Mendelian transmission and the null hypothesis (i.e., no linkage or association).

‡Variance of the test statistic S.

Table XII. Family-based association analysis of promoter SNPs in DNA double-strand break repair genes

Gene and SNP	DNA variant	MAF	No. of families*	S-E(S)†	Var(S)‡	Z	P
<i>ATM</i>							
rs4987876	G>T	.10	48	0.50	13.75	0.14	.89
rs228589	T>A	.38	99	4.00	32.00	0.71	.48
<i>BRCA1</i>							
rs4793204	T>C	.34	114	-5.00	37.00	-0.82	.41
rs799906	A>G	.35	118	-6.50	37.75	-1.06	.29
rs8176071	INS ACA	.35	119	-6.0	38.50	-0.97	.33
rs3092986	A>G	.05	27	-1.50	6.75	-0.58	.56
<i>BRCA2</i>							
rs206114	C>A	.42	126	1.50	40.75	0.24	.81
rs3072036	DEL GTCTAA	.41	114	4.50	38.75	0.72	.47
rs206115	A>G	.44	111	6.00	38.00	0.97	.33
rs206116	C>T	.42	103	3.00	35.50	0.50	.61
rs206117	C>T	.42	112	7.00	38.00	1.14	.26
rs3092989	G>A	.18	90	-9.00	27.50	-1.72	.09
<i>RAD51</i>							
rs2619679	A>T	.50	120	-1.00	40.00	-0.16	.87
<i>XRCC4</i>							
rs3763063	T>C	.44	127	-5.5	42.75	-0.84	.40
rs1993947	C>G	.11	47	3.0	13.50	0.82	.41
<i>XRCC5</i>							
rs828907	G>T	.44	137	0.0	47.00	0.00	1
rs11685387	C>T	.31	107	8.0	36.50	1.32	.19
<i>XRCC6</i>							
rs28384701	C>T	.06	39	3.5	10.25	1.09	.27
rs2267437	C>G	.43	117	1.0	38.50	0.16	.87
rs132770	G>A	.21	106	-0.5	30.25	-0.90	.93

FBAT analyses were performed under the additive model. Only the results for the minor alleles are shown. A positive Z value indicates overtransmission of a given allele to affected offspring; a negative Z value indicates undertransmission. Nominally significant results are shown in bold. MAF indicates minor (i.e., variant) allele frequency.

*Number of informative families (i.e., families with at least one heterozygote parent).

†Test statistic (S) from FBAT for the observed number of transmitted alleles – the expected value of S (E(S)) under Mendelian transmission and the null hypothesis (i.e., no linkage or association).

‡Variance of the test statistic S.

Table XIII. Family-based association analysis of promoter haplotypes in DNA double-strand break repair genes

Gene and Haplotype	Alleles	MAF	No. of families*	S-E(S)†	Var(S)‡	Z	P	Global χ^2 (df)	Global P
ATM									
ATM*1	GT	.55	105	-3.46	34.71	-0.59	.56	1.53 (3)	.67
ATM*2	GA	.36	96	3.46	33.73	0.60	.55		
ATM*3	TT	.08	43	-0.55	13.23	-0.15	.88		
ATM*4	TA	.002	0						
BRCA1									
BRCA1*1	TA(-)A	.63	107	8.00	35	1.35	.18	4.36 (3)	.22
BRCA1*2	CG(ACA)A	.32	95	-5.50	34.25	-0.94	.35		
BRCA1*3	TA(-)G	.04	25	-0.50	6.25	-0.20	.84		
BRCA2									
BRCA2*1	A(GTCTAA)GCCG	.40	90	8.48	38.23	1.37	.17	12.08 (3)	0.007
BRCA2*2	C(-)ATTG	.38	81	6.00	33.50	1.04	.30		
BRCA2*3	C(GTCTAA)ACCA	.16	65	-7.49	22.75	-1.57	.12		
XRCC4									
XRCC4*1	CC	.45	120	-5.08	41.89	-0.79	.43	1.51 (3)	.68
XRCC4*2	TC	.46	121	1.08	41.26	0.17	.87		
XRCC4*3	TG	.09	44	3.92	11.89	1.14	.26		
XRCC5									
XRCC5*1	TC	.43	117	0.50	47.75	0.07	.94	3.95 (2)	.14
XRCC5*2	GC	.27	108	-9.50	33.75	-1.64	.10		
XRCC5*3	GT	.30	87	9.00	33.50	1.56	.12		
XRCC6									
XRCC6*1	CGG	.42	114	2.41	38.14	0.39	.70	3.24 (4)	.52
XRCC6*2	CCG	.39	121	-0.41	40.43	-0.07	.95		
XRCC6*3	CCA	.14	75	-5.52	20.69	-1.21	.23		
XRCC6*4	TCA	.04	38	3.93	8.87	1.32	.19		

Haplotype-specific FBAT analyses were performed under the additive model. A multiallelic chi-square test (χ^2) was also carried out in FBAT to obtain the global haplotype association significance in the family-based setting. A positive Z value indicates overtransmission of a given haplotype to affected offspring; a negative Z value indicates undertransmission. Nominally significant results are shown in bold. Freq. indicates haplotype frequency; df indicates degrees of freedom; —, not applicable due to lack of informative families (i.e., < 10).

*Number of informative families (i.e., families with at least 1 heterozygote parent).

- †Test statistic (S) from FBAT for the observed number of transmitted haplotypes
– expected value of S ($E(S)$) under Mendelian transmission and the null hypothesis (i.e., no linkage or association).
- ‡Variance of the test statistic S .

REFERENCES

1. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*. 2003 Jan;19(1):149-50.
2. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. PBAT: tools for family-based association studies. *Am J Hum Genet*. 2004 Feb;74(2):367-9.
3. Gauderman W, Morrison J. QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, <http://hydra.usc.edu/gxe>. 2006.
4. Becker T, Knapp M. Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet Epidemiol*. 2004 Jul;27(1):21-32.
5. Herold C, Becker T. Genetic association analysis with FAMHAP: a major program update. *Bioinformatics*. 2009 Jan 1;25(1):134-6.
6. Horvath S, Xu X, Laird NM. The family based association test method: strategies for studying general genotype--phenotype associations. *Eur J Hum Genet*. 2001 Apr;9(4):301-6.
7. Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. *Genet Epidemiol*. 2000;19 Suppl 1:S36-42.

APPENDIX III

*RESULTS OF THE FETOMATERNAL
ASSOCIATION TESTS FOR THE DNA
DOUBLE-STRAND BREAK REPAIR GENES*

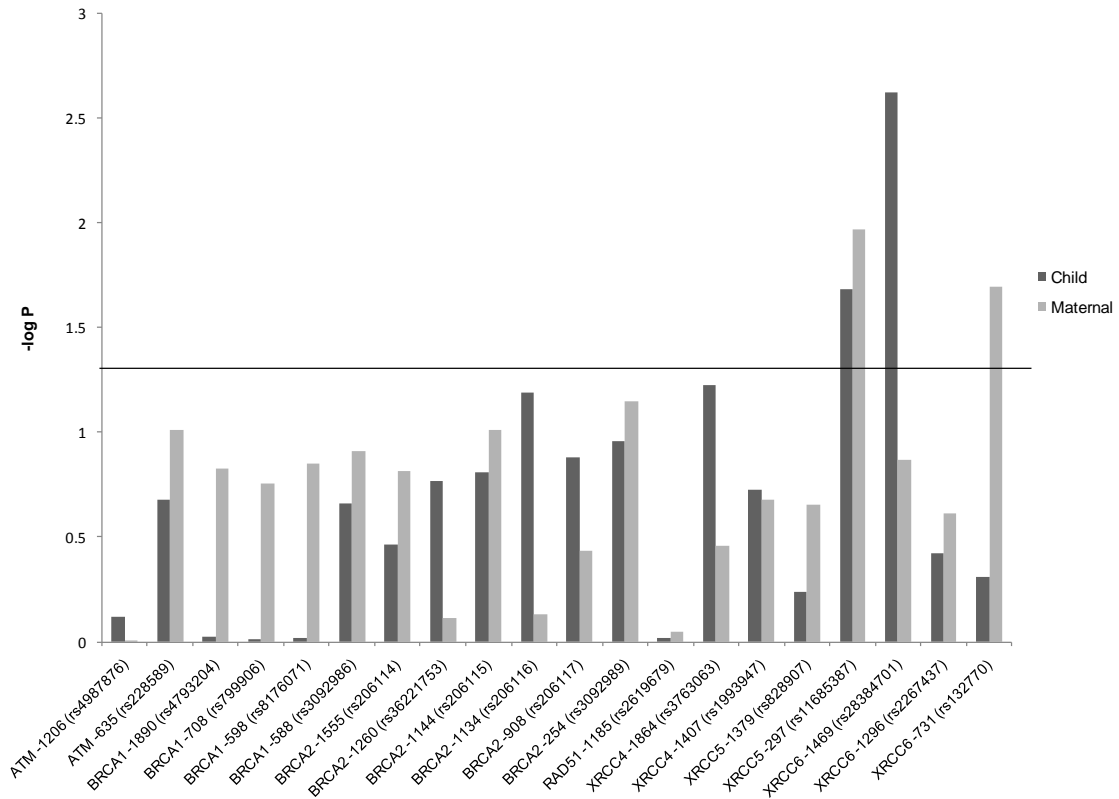


Figure I. Log-linear, likelihood-ratio association analysis between 20 regulatory SNPs from 7 DNA double-strand break repair genes and childhood pre-B acute lymphoblastic leukemia

Log-linear regression analysis was performed in LEM (1-2) using 203 case-triads, 118 unrelated ALL patients and 329 controls. Results of the likelihood-ratio chi-square tests ($-\log P$) are shown for the single-step Child and Mother tests. A three degree-of-freedom likelihood-ratio test for mating asymmetry (MA) was performed using a P value < 0.10 to reject symmetry. Under this threshold, mating symmetry (i.e. six mating-type parameters) was assumed at all loci but variants rs36221753 and rs206116 of the *BRCA2* gene for which MA models (nine mating-type parameters) were used to test for association. Horizontal reference line indicates P value of 0.05.

Table I. Log-linear, likelihood-ratio association analysis between 20 regulatory SNPs from 7 double-strand break repair genes and childhood pre-B ALL

Gene, DNA variant	Model	LR Chi2 (df)	P
<i>ATM</i>			
rs4987876	Child vs. Null	0.5477 (2)	0.7604
	Mother vs. Null	0.0126 (2)	0.9937
rs228589	Child vs. Null	3.1310 (2)	0.2090
	Mother vs. Null	4.6616 (2)	0.0972
<i>BRCA1</i>			
rs4793204	Child vs. Null	0.1228 (2)	0.9405
	Mother vs. Null	3.7985 (2)	0.1497
rs799906	Child vs. Null	0.0573 (2)	0.9717
	Mother vs. Null	3.4793 (2)	0.1756
rs8176071	Child vs. Null	0.0826 (2)	0.9595
	Mother vs. Null	3.9198 (2)	0.1409
rs3092986	Child vs. Null	3.0368 (2)	0.2191
	Mother vs. Null	4.1961 (2)	0.1227
<i>BRCA2</i>			
rs206114	Child vs. Null	2.1435 (2)	0.3424
	Mother vs. Null	3.7607 (2)	0.1525
rs36221753	Child vs. Null	3.5219 (2)	0.1719
	Mother vs. Null	0.5142 (2)	0.7733
rs206115	Child vs. Null	3.7406 (2)	0.1541
	Mother vs. Null	4.6467 (2)	0.0979
rs206116	Child vs. Null	5.4814 (2)	0.0645
	Mother vs. Null	0.6138 (2)	0.7357
rs206117	Child vs. Null	4.0435 (2)	0.1324
	Mother vs. Null	2.0003 (2)	0.3678

rs3092989	Child vs. Null	4.4101 (2)	0.1102
	Mother vs. Null	5.2825 (2)	0.0713
RAD51			
rs2619679	Child vs. Null	0.0930 (2)	0.9546
	Mother vs. Null	0.2173 (2)	0.8970
XRCC4			
rs3763063	Child vs. Null	5.6530 (2)	0.0592
	Mother vs. Null	2.1153 (2)	0.3473
rs1993947	Child vs. Null	3.3329 (2)	0.1889
	Mother vs. Null	3.1122 (2)	0.2110
XRCC5			
rs828907	Child vs. Null	1.0927 (2)	0.5791
	Mother vs. Null	3.0183 (2)	0.2211
rs11685387	Child vs. Null	7.7518 (2)	0.0207
	Mother vs. Null	9.0817 (2)	0.0107
	Child + Mother vs. Null	13.8334 (4)	0.0078
	Child + Mother vs. Mother	4.7517 (2)	0.0929
	Child + Mother vs. Child	6.0815 (2)	0.0478
XRCC6			
rs28384701	Child vs. Null	12.0886 (2)	0.0024
	Mother vs. Null	4.0077 (2)	0.1348
	Child + Mother vs. Null	13.2668 (4)	0.0100
	Child + Mother vs. Child	1.1782 (2)	0.5548
rs2267437	Child vs. Null	1.9515 (2)	0.3769
	Mother vs. Null	2.8099 (2)	0.2454
rs132770	Child vs. Null	1.4259 (2)	0.4902
	Mother vs. Null	7.8070 (2)	0.0202
	Child + Mother vs. Null	7.8453 (4)	0.0974
	Child + Mother vs. Mother	0.0383 (2)	0.9810

Likelihood-ratio tests were performed in a forward stepwise fashion. The most significant single-step test (Child vs. Null or Mother vs. Null) was tested against a joint effects model in a 2 degree of freedom likelihood-ratio test (Child + Mother vs. Child or Child + Mother vs. Mother). A three degree-of-freedom likelihood-ratio test for mating asymmetry (MA) was performed using a P value < 0.10 to reject symmetry. Under this threshold, mating symmetry (i.e. six mating-type parameters) was assumed at all loci but variants rs36221753 and rs206116

of the *BRCA2* gene for which MA models (nine mating-type parameters) were used to test for association. Nominally significant P values are shown in bold remain. LR Chi2 indicates likelihood-ratio chi-square test; df, degrees of freedom.

REFERENCES

1. van Den Oord EJ, Vermunt JK. Testing for linkage disequilibrium, maternal effects, and imprinting with (In)complete case-parent triads, by use of the computer program LEM. *Am J Hum Genet.* 2000 Jan;66(1):335-8.
2. Weinberg CR, Umbach DM. A hybrid design for studying genetic influences on risk of diseases with onset early in life. *Am J Hum Genet.* 2005 Oct;77(4):627-36.

APPENDIX IV

*FOLLOW-UP ANALYSIS OF A GENOME-WIDE
ASSOCIATION STUDY IDENTIFIES CDKN2A AS A
SUSCEPTIBILITY LOCUS FOR CHILDHOOD ACUTE
LYMPHOBLASTIC LEUKEMIA*

This research was originally published in *Nature Genetics*. 2010;42(6):492-4.
 © 2010 Nature Publishing Group, a division of Macmillan Publishers Limited.

BRIEF COMMUNICATIONS

nature
genetics

Variation in *CDKN2A* at 9p21.3 influences childhood acute lymphoblastic leukemia risk

Amy L. Sherborne^{1,20}, Fay J. Hosking^{1,20}, Rashmi B. Prasad², Rajiv Kumar², Rolf Koehler³, Jayaram Vijayakrishnan¹, Elli Papaemmanuil¹, Claus R. Bartram³, Martin Stanulla⁴, Martin Schrappe⁴, Andreas Gast², Sara E. Dobbins¹, Yussanne Ma¹, Eamonn Sheridan⁵, Malcolm Taylor⁶, Sally E. Kinsey⁷, Tracey Lightfoot⁸, Eve Roman⁸, Julie A. E. Irving⁹, James M. Allan⁹, Anthony V. Moorman⁹, Christine J. Harrison⁹, Ian P. Tomlinson¹⁰, Sue Richards¹¹, Martin Zimmermann¹², Csaba Szalai¹³, Ágnes F. Semsei¹³, Daniel J. Erdelyi¹³, Maja Krajcinovic¹⁴, Daniel Sinnott¹⁴, Jasmine Healy¹⁴, Anna Gonzalez Neira¹⁵, Norihiko Kawamata¹⁶, Seishi Ogawa¹⁷, H. Phillip Koeffler¹⁸, Karl Hemminki², Mel Greaves¹⁹ & Richard S. Houlston^{1,20}

Using data from a genome-wide association study of 907 individuals with childhood acute lymphoblastic leukemia (cases) and 2,398 controls and with validation in samples totaling 2,386 cases and 2,419 controls, we have shown that common variation at 9p21.3 (rs3731217, intron 1 of *CDKN2A*) influences acute lymphoblastic leukemia risk (odds ratio = 0.71, $P = 3.01 \times 10^{-11}$), irrespective of cell lineage.

Acute lymphoblastic leukemia (ALL) is the major pediatric cancer in western countries, with B-cell precursor (BCP) ALL accounting for ~70% of cases. Although there is little evidence for a strong familial basis to ALL, evidence for an inherited susceptibility has recently been provided by genome-wide association (GWA) studies showing that *IKZF1* (7p12.2), *ARID5B* (10q21.2) and *CEBPE* (14q11.2) variants confer a modest but significant risk^{1,2}.

Our previously reported genome-wide association (GWA) studies of ALL were based on pooling data from two case-control series comprising a total of 907 cases and 2,398 controls successfully genotyped for 291,371 tagging SNPs¹. To search for additional new variants

influencing the risk of ALL, we have performed replication of 34 SNPs selected on the basis of statistical significance ($P < 0.0001$) coupled with considerations of minor allele frequency > 0.05 , Hardy-Weinberg equilibrium ($P > 0.05$ in controls) and potential candidacy of nearby genes (on the basis of a role in B-cell and/or cancer biology) (Supplementary Table 1) in additional case-control series. The 34 SNPs were genotyped in an independent series of 1,428 ALL cases ascertained through the German Berlin-Frankfurt-Munster (BFM) group childhood ALL trials and 1,516 population controls. Thirty-three SNPs showed no significant evidence of an association at $P_{\text{trend}} < 0.01$ in this series and were not further evaluated (Supplementary Table 1). rs3731217, mapping to 9p21.3, provided strong evidence for an association in the German series ($P = 1.15 \times 10^{-7}$) and was genotyped in case-control series from Spain (148 cases and 187 controls), Hungary (550 cases and 450 controls) and Canada (260 cases and 266 controls) (details in Supplementary Methods). Jointly, these four replication series provided strong evidence for association (combined odds ratio (OR) = 0.68, $P = 1.13 \times 10^{-8}$). Pooling genotype data for all of the studies provided unequivocal evidence for a relationship between rs3731217 and ALL risk (statistically significant after adjustment for multiple testing, assuming a conservative Bonferroni correction for the 291,371 tests in the original GWA study; OR for the T allele = 0.71, 95% CI 0.64–0.78, $P = 3.01 \times 10^{-11}$) (Fig. 1). This association was consistent in each of the replication series with nonsignificant statistics of between-study heterogeneity ($P = 0.76$, $I^2 = 0.0\%$).

rs3731217 localizes to intron 1 of *CDKN2A* (located at 21,974,661 base pairs; Fig. 2 and Supplementary Fig. 1) within a 174-kb region of linkage disequilibrium (LD) at 9p21.3. This region encompasses the *CDKN2A* and *CDKN2B* tumor-suppressor genes and the noncoding antisense RNA encoded by *CDKN2BAS*. *CDKN2A* encodes both p16 (INK4A), a negative regulator of cyclin-dependent kinases, and p14 (ARF1), an activator of p53. *CDKN2A* and *CDKN2B* are frequently inactivated in multiple hematological malignancies, and mono- or biallelic deletion of *CDKN2A* is one of the most frequent genetic events in childhood B- and T-lineage ALL³. *CDKN2A* deletions arise as secondary genetic events in cases of ALL initiated by *ETV6-RUNX1* gene fusions⁴ and increase in frequency in cases of ALL relapse. Recent GWA studies have identified variation at 9p21.3 to be associated with

¹Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey, UK. ²Division of Molecular Genetic Epidemiology, German Cancer Research Centre, Heidelberg, Germany. ³Institute of Human Genetics, University of Heidelberg, Heidelberg, Germany. ⁴Department of Pediatrics, Acute Lymphoblastic Leukemia-Berlin-Frankfurt-Munster (ALL-BFM) Clinical Trial Center, University of Kiel, Kiel, Germany. ⁵Yorkshire Regional Genetic Service, St. James's University Hospital, Leeds, UK. ⁶Cancer Immunogenetics Group, School of Cancer Sciences, University of Manchester, Research Floor, St. Mary's Hospital, Manchester, UK. ⁷Department of Paediatric and Adolescent Oncology and Haematology, St. James's University Hospital, Leeds, UK. ⁸Epidemiology and Genetics Unit, Department of Health Sciences, University of York, York, UK. ⁹Northern Institute for Cancer Research, Newcastle University, Newcastle upon Tyne, UK. ¹⁰Molecular and Population Genetics, Wellcome Trust Centre for Human Genetics, Oxford, UK. ¹¹Clinical Trial Service Unit and Epidemiological Studies Unit, Richard Doll Building, Old Road Campus, Roosevelt Drive, Oxford, UK. ¹²Department of Pediatric Hematology and Oncology, Medical School, Hannover, Germany. ¹³Department of Genetic, Cell- and Immunobiology, SNP Core Facility Laboratorium, Semmelweis University, Budapest, Hungary. ¹⁴Division of Hematology-Oncology, Research Center of the Sainte-Justine University Health Center, University of Montréal, Montréal, Québec, Canada. ¹⁵Spanish National Cancer Research Center (CNIO), Madrid, Spain. ¹⁶Division of Hematology and Oncology, Cedars-Sinai Medical Center, University of California Los Angeles School of Medicine, Los Angeles, California, USA. ¹⁷Regeneration Medicine of Hematopoiesis, University of Tokyo, School of Medicine, Tokyo, Japan. ¹⁸National University of Singapore Institute of Human Genetics, Singapore, Singapore. ¹⁹Section of Haemato-oncology, Institute of Cancer Research, Sutton, Surrey, UK. ²⁰These authors contributed equally to this work. Correspondence should be addressed to R.S.H.

Received 25 January; accepted 12 February; published online 9 May 2010; doi:10.1038/ng.585



BRIEF COMMUNICATIONS

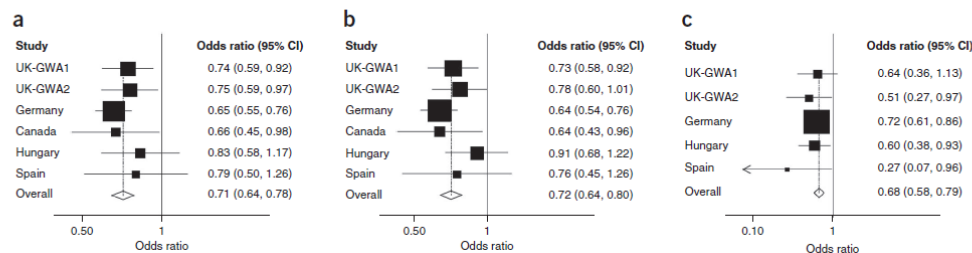


Figure 1 Forest plots of effect size and direction for 9p21.3 (rs3731217) association. (a–c) Association between all cases of ALL (a), BCP-ALL cases (b) and T-ALL cases and controls (c). Boxes denote OR point estimates, with their areas being proportional to the inverse variance weight of the estimate. Horizontal lines represent 95% CIs. The diamond and dashed line represent the summary OR computed under a fixed effects model, with the 95% CI indicated by the width of the diamond. The unbroken vertical line is at the null value (OR = 1.0).

nevus density⁵, melanoma⁶, basal cell carcinoma⁷, glioma⁸, type 2 diabetes⁹ and coronary heart disease¹⁰. It is possible that these associations are independent, but given the extent of LD across the region, it is also plausible that these multiple diseases have a common causal basis.

Given the biological heterogeneity of ALL, we analyzed the association between the major subtypes of ALL and rs3731217 genotype. The association remained highly significant when the analysis was confined to BCP-ALL (combined OR = 0.72, 95% CI 0.64–0.80, $P = 5.29 \times 10^{-10}$, P of heterogeneity (P_{het}) = 0.41, I^2 = 1.8%) and also for T-cell ALL (combined OR = 0.68, 95% CI 0.58–0.79, $P = 1.88 \times 10^{-7}$, P_{het} = 0.46, I^2 = 0.0%) (Fig. 1). Subtype analysis of BCP-ALL in the UK GWA studies and the German case-control series provided no strong evidence that variation at 9p21.3 is associated with risk of a specific cytogenetic type of BCP-ALL ($P = 0.22$). Analysis of data from 378 UK-GWA2 study and 1,302 German cases provided no evidence that rs3731217 genotype was associated with event-free survival ($P = 0.8$ and $P = 0.15$ respectively; Supplementary Fig. 2). This is perhaps not unexpected given that *CDKN2A-CDKN2B* deletion status itself is not consistently associated with outcome in childhood ALL¹¹.

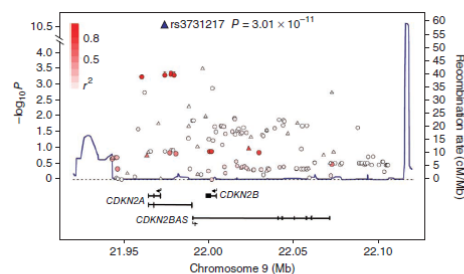
Although *CDKN2A* is a strong candidate ALL-susceptibility gene on the basis of its biology, we cannot exclude a role of the other genes in 9p21.3 region as the basis of the association. To explore the entire genetic interval, we imputed 320 SNP genotypes within the 174-kb LD block containing rs3731217, using both HapMap and 1000 Genomes data in our GWA series (at 21,942,000–22,116,000 base pairs). Five SNPs showed a $P_{trend} < 10^{-3}$, and all were located within 15 kb of rs3731217 (Supplementary Table 2). None of these SNPs provided substantially better evidence for an association than rs3731217, and all were in strong LD with rs3731217 ($r^2 > 0.80$, $D' > 0.97$). Results from haplotype analysis provided evidence for a single risk variant

defined by rs3731239-rs2811709-rs4074785-rs3731217-rs2811712-rs3218018 (Supplementary Table 3). To gain further insight into the causal basis of the association, we performed logistic regression using SNPs mapping to the 174-kb region of LD at 9p21.3. Analyses conditional on rs3731217 provided no support for a secondary association; however, a single disease-causing haplotype as inferred by ancestral recombination graph analysis could not be fully defined (Supplementary Table 3).

Inactivation of *CDKN2A* and *CDKN2B* is primarily a consequence of mono- or biallelic 9p21.3 deletion rather than promoter methylation^{11,12}. To explore the possibility that the risk allele for rs3731217 is favored by somatic tumor evolution, we examined the relationship between genotype and *CDKN2A* and *CDKN2B* deletion in leukemic clones in 66 UK-GWA2 and 387 German cases. Although it is mechanistically plausible that somatic loss of the wild-type allele preferentially occurs in the leukemic clones of carriers, the frequency of *CDKN2A* deletion (mono- or biallelic) in leukemic clones was not different in carriers of the risk allele compared with non-carriers ($P = 0.07$ and $P = 0.93$ in the two datasets, respectively; Supplementary Table 4).

HapMap and 1000 Genomes data provided no evidence that rs3731217 is strongly correlated with a coding SNP in *CDKN2A*, *CDKN2B* or *CDKN2BAS* ($r^2 < 0.001$, $D' < 0.001$; Supplementary Table 5). To explore the possibility that the association to disease might be mediated through differential *CDKN2A* or *CDKN2B* expression, we investigated the relationship between rs3731217 genotype and mRNA expression in 90 Epstein-Barr virus lymphoblastoid cells¹³. Although we found no association between rs3731217 genotype and expression of either mRNA transcript (Supplementary Fig. 3), steady-state levels of RNA at a single time point and in cycling mature B-cells may not adequately capture the impact of differential expression in leukemogenesis.

Figure 2 LD structure and association results for the 9p21.3 ALL locus. The local recombination rate is plotted in dark blue over this 174-kb chromosomal segment. Armitage trend test P values (as $-\log_{10} P$ values along the left y axis; note the broken axis between 4.0 and 10.5) are shown for the SNPs analyzed. Each triangle represents a SNP genotyped in the GWA study, circles represent imputed SNPs, and the most associated SNP in the combined analysis, rs3731217, is marked by a triangle (blue in combined analysis). The color intensity of each \square reflects the extent of LD with rs3731217, indicated by red ($r^2 > 0.8$) going to white ($r^2 < 0.2$). Physical positions are based on NCBI build 36 of the human genome. Also shown are the relative positions of genes mapping to each region of association. Exons of genes have been redrawn to show the relative positions in the gene, and therefore maps are not to physical scale.



BRIEF COMMUNICATIONS

Elucidation of the causal basis of the 9p21.3 association will be contingent on fine-mapping and functional studies. To examine if any directly typed or imputed SNPs annotate a predicted transcription factor binding or enhancer elements, we conducted a bioinformatic search of the 174-kb region using TRANSFAC matrix database, PReMod and EEL software (see URLs). None of the SNPs associated with ALL risk (that is, having, $P < 0.001$) mapped within predicted regulatory elements (Supplementary Table 6 and Supplementary Fig. 1).

Irrespective of the causal basis of the 9p21.3 association, the potential impact of common alleles on gene expression will be modest and could occur at any time before disease diagnosis. Moreover, expression differences may only be relevant to a subpopulation of cells that provide 'targets' for leukemogenic mutations. There is some evidence that *CDKN2A* and *CDKN2B* deletions might be initiated by 'off-target' effects of the lymphoid mutagenic enzymes recombination activating proteins 1 and 2 (RAG1 and RAG2)¹⁴ or activation-induced cytidine deaminase (AID)¹⁵. Therefore, another possibility is that the association between increased risk of ALL and inherited variation in *CDKN2A* and *CDKN2B* reflects its structural or sequence-based vulnerability as a substrate.

There was no evidence of significant interaction between rs3731217 and the previously identified^{1,2} risk loci at 7p12.2 (*IKZF1*, rs4132601), 10q21.2 (*ARID5B*, rs7089424) and 14q11.2 (*CEBPE*, rs2239633), an observation compatible with each locus having an independent effect on ALL risk. Although the risk of ALL associated with these loci is modest, collectively they play a substantial role in the development of ALL, jointly accounting for as much as ~80% of the attributable risk in European populations. It will be intriguing to examine whether our new findings translate to nonwestern populations that have a different prevalence of ALL.

In summary, we have identified a new ALL risk locus at 9p21.3, and these findings provide additional insight into the development of ALL. Further studies are required to identify the causal variant(s) and to elucidate the biological basis of the association between genetic variation at *CDKN2A* loci and ALL pathogenesis.

URLs. R suite, <http://www.r-project.org/>; detailed information on the tag SNP panel, <http://www.illumina.com/>; dbSNP, <http://www.ncbi.nlm.nih.gov/projects/SNP/>; HapMap, <http://www.hapmap.org/>; 1958 Birth Cohort, <http://www.cls.ioe.ac.uk/studies.asp?section=000100020003>; MRC ALL 97 (Protocol 97PRT/14), <http://www.thelancet.com/protocol-reviews/97PRT-14>; United Kingdom Childhood Cancer Study, <http://www.ukccs.org/>; KBioscience, <http://kbioscience.co.uk/>; WGAViewer, <http://www.genome.duke.edu/centers/pg2/downloads/wgaviewer.php>; IMPUTE, <http://mathgen.stats.ox.ac.uk/impute/impute.html>; SNPTEST, <http://www.stats.ox.ac.uk/~marchini/software/gwas/snpctest.html>; Margarita, <http://www.sanger.ac.uk/resources/software/margarita/>; EEL, <http://www.cs.helsinki.fi/u/kpalin/EEL/>; PReMod, <http://genomequebec.mcgill.ca/PReMod/welcome.do>; TRANSFAC matrix database, <http://www.biobase-international.com/pages/index.php?id=transfac>; 1000 Genomes, <http://www.1000genomes.org>; JASPAR2 database, <http://jaspar.cgb.ki.se/>; collaborators to the 1958 Birth Cohort dataset, <http://www.wtccc.org.uk/>.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The Kay Kendall Leukemia Fund and Leukemia Research (UK) provided principal funding for the study. Support from Cancer Research UK (C1298/A8362, supported by the Bobby Moore Fund) is also acknowledged. The study made use of genotyping data from the 1958 Birth Cohort which was generated and generously supplied to us by Panagiotis Deloukas of the Wellcome Trust Sanger

Institute. For a full list of the investigators who contributed to the generation of the 1958 data, see URL section. We are grateful to S. Richards and J. Burrett (Clinical Trials Service Unit, Oxford) and L. Chilton (Leukaemia Research Cytogenetics Group, Northern Institute of Cancer Research, Newcastle Univ.), J. Simpson (Univ. York), P. Thomson and A. Hussain (Cancer Immunogenetics, School of Cancer Sciences, Univ. Manchester) for assistance with data harmonization. We thank the Leukaemia Research (LR) Childhood Cancer Leukaemia Group (CCLG) Childhood Leukaemia Cell Bank for access to the Medical Research Council ALL trial samples. We are grateful to the UK Cancer Cytogenetics Group (UKCCG) for data collection and provision of samples. P. Thomson (Cancer Immunogenetics, School of Cancer Sciences, University of Manchester) is funded by Children with Leukemia and we acknowledge their support. Genotyping of German cases and controls was partly covered by funding from Tumorzentrum Heidelberg-Mannheim, EU (Food-CT-2005-016320 and Health-F3-2007-200767) and Deutsche Krebshilfe. The Hungarian sample collection was financially supported by grants from the Hungarian Scientific Research Fund (T042500), the Economic Competitiveness Operational Programme, Hungary (GVOP 3.1.1-2004-05-0022/3.0) and NKTH (National Research and Technology) TECH_08-A1/2-2008-0120. M.K. is a scholar of the Fonds de la Recherche en Santé du Québec. D.S. holds the François-Karl Viau Chair in Pediatric Oncogenomics and is also a scholar of the Fonds de la Recherche en Santé du Québec. J.H. is the recipient of a National Sciences and Engineering Research Council Canada Graduate's scholarship. P. Garcia-Miguel and A. Sastre from Hospital Infantil La Paz, Madrid, and J. Luis Vivanco Martínez from Hospital 12 de Octubre, Madrid, were involved in the sample collection of the Spanish samples, and R. Alonso from Centro Nacional de Investigaciones Oncológicas was involved in the genotyping.

We are grateful to all the study subjects and individuals for their participation and, finally, we would also like to thank the clinicians, other hospital staff and study staff who contributed to the blood sample and data collection for this study.

AUTHOR CONTRIBUTIONS

R.S.H. and M.G. obtained financial support. R.S.H. designed and provided overall project management. R.S.H. drafted the manuscript with contributions from E.J.H., A.L.S. and M.G.; A.L.S. performed overall project management, development, database development and oversaw laboratory analyses; E.J.H. performed statistical analyses; E.J.H. and A.L.S. performed bioinformatics analyses; J.V. and E.P. performed UK sample preparation and genotyping; E.S. and S.E.K. performed curation and sample preparation of the Medical Research Council ALL 97 trial samples; T.L. and E.R. managed and maintained UKCCS sample data; M.T. performed curation and sample preparation of United Kingdom Childhood Cancer Study samples; J.M.A. and J.A.E.L. performed ascertainment, curation and sample preparation of the Northern Institute for Cancer Research case series. I.P.T. generated and managed UK colorectal cancer control genotypes. A.V.M. and C.J.H. performed UK *CDKN2A* deletion analysis; N.K., S.O. and H.P.K. carried out German *CDKN2A* deletion analysis; S.E.D. and Y.M. carried out HapMap and 1000 Genomes imputation; S.R. carried out survival analysis of UK data; M.Z. carried out survival analysis of German data; K.H. oversaw analysis of the German cohort; R.B.P., A.G. and R.K. conducted genotyping of German samples; R. Koehler, M. Stanulla, M. Schrappe and C.R.B. provided German DNA for analysis; D.J.E. and C.S. coordinated the data and sample collection of the Hungarian ALL cohort; A.E.S. genotyped the Hungarian samples; M.K., D.S. and J.H. performed curation and preparation of Canadian samples; A.G.N. was responsible for curation, management and genotyping of Spanish samples. All authors contributed to the final paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturegenetics/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

1. Papaemmanuil, E. *et al.* *Nat. Genet.* **41**, 1006–1010 (2009).
2. Treviño, L.R. *et al.* *Nat. Genet.* **41**, 1001–1005 (2009).
3. Mullighan, C.G. & Downing, J.R. *Leukemia* **23**, 1209–1218 (2009).
4. Bateman, C. *et al.* *Blood* published online. doi:10.1182/blood-2009-10-25143 (8 January 2010).
5. Falchi, M. *et al.* *Nat. Genet.* **41**, 915–919 (2009).
6. Bishop, D.T. *et al.* *Nat. Genet.* **41**, 920–925 (2009).
7. Stacey, S.N. *et al.* *Nat. Genet.* **41**, 909–914 (2009).
8. Shete, S. *et al.* *Nat. Genet.* **41**, 899–904 (2009).
9. Scott, L.J. *et al.* *Science* **316**, 1341–1345 (2007).
10. McPherson, R. *et al.* *Science* **316**, 1488–1491 (2007).
11. Suijng, S. *et al.* *Blood* **113**, 100–107 (2009).
12. Mirebeau, D. *et al.* *Haematologica* **91**, 881–885 (2006).
13. Stranger, B.E. *et al.* *PLoS Genet.* **1**, e78 (2005).
14. Kitagawa, Y. *et al.* *J. Biol. Chem.* **277**, 46289–46297 (2002).
15. Iacobucci, I. *et al.* *Leukemia* **24**, 66–73 (2010).

Reprint Permissions

NATURE PUBLISHING GROUP LICENSE TERMS AND CONDITIONS

May 19, 2010

This is a License Agreement between Jasmine Healy ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2432700720080
License date	May 19, 2010
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Genetics
Licensed content title	Variation in <i>CDKN2A</i> at 9p21.3 influences childhood acute lymphoblastic leukemia risk
Licensed content author	Amy L Sherborne, Fay J Hosking, Rashmi B Prasad, Rajiv Kumar, Rolf Koehler, Jayaram Vijayakrishnan, Elli Papaemmanuil, Claus R Bartram, Martin Stanulla, Martin Schrappe, Andreas Gast, Sara E Dobbins, Yussanne Ma, Eamonn Sheridan, Malcolm Taylor, Sally E Kinsey, Tracey Lightfoot, Eve Roman, Julie A E Irving, James M Allan, Anthony V Moorman, Christine J Harrison, Ian P Tomlinson, Sue Richards, Martin Zimmermann, Csaba Szalai, Ágnes F Semsei, Daniel J Erdelyi, Maja Krajcinovic, Daniel Sinnett, Jasmine et al.
Volume number	
Issue number	
Pages	
Year of publication	2010
Portion used	Full paper
Requestor type	Student
Type of Use	Thesis / Dissertation
Billing Type	Invoice
Company	Jasmine Healy
Billing Address	3175 chemin de la Côte Sainte-Catherine Room B-490 Montreal, QC H3T 1C5 Canada
Customer reference info	
Total	0.00 USD
Terms and Conditions	Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use, subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.
2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version. Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run). NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.
3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).
4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.
5. The credit line should read:

Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME]
(reference citation), copyright (year of publication)

For AOP papers, the credit line should read:

Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME],
advance online publication, day month year (doi: 10.1038/sj.[JOURNAL
ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME]
(reference citation), copyright (year of publication)
7. Translations of 401 words up to a whole article require NPG approval. Please visit <http://www.macmillanmedicalcommunications.com> for more information. Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME]
(reference citation), copyright (year of publication).

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.

v1.1

Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK10786864.

Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

**Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006**

If you find copyrighted material related to this license will not be used and wish to cancel, please contact us referencing this license number 2432700720080 and noting the reason for cancellation.

Questions? customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

CURRICULUM VITAE

Jasmine Healy

Ph.D. Candidate, University of Montreal

ACADEMICS

2006 – 2010	Ph.D. in Molecular Biology University of Montreal and the Sainte-Justine Hospital Research Centre, <i>Montreal, Qc</i> Gene-gene interactions and parental genetic effects in the susceptibility to childhood leukemia Advisor: Daniel Sinnett
2005 – 2006	M.Sc. in Genetic Epidemiology University of Sheffield, <i>Sheffield, UK</i> Role of genetic determinants and gene-gene interactions in the susceptibility and severity of rheumatoid arthritis Advisors: Kevin Walters and Gerry Wilson
2003 – 2005	M.Sc. in Molecular Biology University of Montreal and the Sainte-Justine Hospital Research Centre, <i>Montreal, Qc</i> Regulatory Polymorphisms in Genes Involved in Cell Cycle Control and the Susceptibility to Childhood Leukemia Advisor: Daniel Sinnett
2000 – 2003	B.Sc. with Honours in Biopharmaceutical Sciences University of Ottawa, <i>Ottawa, On</i> Functional Characterization of Two Novel Proteins, LC27 and ANKT, and their Putative Interaction with ATRX Advisor: David Picketts
1998 – 2000	D.E.C. in Health Sciences Collège Jean-de-Brébeuf, <i>Montreal, Qc</i>

BURSARIES, AWARDS AND DISTINCTIONS

2008 – 2011	Alexander Graham Bell Canada Graduate Scholarship (CGS-D3), \$105,000 Natural Sciences and Engineering Research Council of Canada (NSERC)
2007 – 2010	Doctoral Postgraduate Scholarship (PGS-D3), \$63,000 Natural Sciences and Engineering Research Council of Canada (NSERC) Accepted until January 2008

2007 – 2010	Bourse de formation de doctorat, \$60,000 Fonds de la Recherche en Santé du Québec (FRSQ) Declined
2007 – 2008	Doctoral Research Bursary, \$28,000 Cole Foundation – Accepted until March 2007
2006 – 2007	Doctoral Research Bursary, \$15,000 Foundation for Research into Children's Diseases and the Sainte-Justine Hospital Foundation
2005	Centenary Scholarship, \$2,167 University of Sheffield
2004 – 2005	Canada Graduate Scholarships – Master's Award, \$17,500 Canadian Institutes of Health Research (CIHR)
2009	Research Excellence Award, \$1,000 Foundation of Stars for Children's Health Research, the 27 ^e Ball and Award Gala of the Foundation of Stars
2007	Student Travel Award, \$750 American Association for Cancer Research (AACR) for the Special Conference on Approaches to Complex Pathways in Molecular Epidemiology
2007	Student Travel Award, \$1,500 National Cancer Institute of Canada (NCI) for the Special Conference on Approaches to Complex Pathways in Molecular Epidemiology
2007	Student Travel Award, \$600 Genome Canada for the Human Genome Organization's 12 th Human Genome Meeting
2007	Doctoral Research Excellence Prize, \$3,000 Sainte-Justine Hospital Foundation
2007	Honorary award for important contributions to the academic and student life Sainte-Justine Hospital Research Centre
2006	Graduated First Class University of Sheffield
2003	Graduated Magna cum laude University of Ottawa
2000	Admissions Bursary, \$2,500 Faculty of Science, University of Ottawa

PUBLICATIONS

1. **Healy J**, Richer C, Dionne J, Bélanger H, Larivière M, Benhamza K, Ouimet M, Gagné M, Weth V, Beaulieu P, Massé H, Bourgey M, Roy-Gagnon MH, Sinnett D. Promoter Variants in Genes Involved in the Cell Cycle and DNA Repair Pathways and the Susceptibility to Childhood Acute Lymphoblastic Leukemia. *Cancer Res.* 2010 June. (Submitted)
2. Sherborne A, Hosking FJ, Prasad RB, Kumar R, Koehler R, Vijayakrishnan J, Papaemmanuil E, Bartram CR, Stanulla M, Schrappe M, Gast A, Sheridan E, Taylor M, Kinsey SE, Lightfoot T, Roman E, Irving JAE, Allan JM, Moorman AV, Harrison CJ, Tomlinson IP, Szalai C, Semsei AF, Erdelyi DJ, Krajcinovic M, Sinnett D, **Healy J**, Neira AG, Hemminki K, Greaves M, Houlston RS. Variation at 9p21.3 (*CKN2A*) influences childhood acute lymphoblastic leukemia risk. *Nat Gen.* 2010;42(6):492-4.
3. **Healy J**, Bourgey M, Richer C, Sinnett D, Roy-Gagnon MH. Detection of fetomaternal genotype associations in early-onset disorders: evaluation of different methods and their application to childhood leukemia. *J Biomed Biotechnol.* 2010; doi:10.1155/2010/369534. (Invited article)
4. **Healy J**, Richer C, Bourgey M, Kritikou EA, Sinnett D. Replication analysis confirms the association of *ARID5B* with childhood B-cell acute lymphoblastic leukemia. *Haematologica.* 2010 May. (Epub ahead of print, doi:10.3324/haematol.2010.022459).
5. **Healy J**, Roy-Gagon MH, Sinnett D. No evidence for association between TGFB1 promoter SNPs and the risk of childhood pre-B acute lymphoblastic leukemia among French Canadians. *Haematologica.* 2009 Jul; 94(7):1034-5.
6. **Healy J**, Dionne J, Bélanger H, Larivière M, Beaulieu P, Labuda D, Sinnett D. Functional impact of sequence variation in the promoter region of TGFB1. *Int. J. Cancer.* 2009 Sept; 125(6):1483-9.
7. Boily G, Beaulieu P, **Healy J**, Sinnett D. Connections Between ETV6-Modulated Genes: Identification of Shared Features. *Cancer Inform.* 2008;6:183-201.
8. **Healy J**, Bérubé NG, Medina CF, Wu S, Hodgson T, Jagla M, Picketts DJ. Patient mutations alter ATRX targeting to PML nuclear bodies. *Eur J Hum Genet.* 2008 Feb;16(2):192-201.

9. **Healy J**, Bélanger H, Beaulieu P, Larivière M, Labuda D, Sinnett D. Promoter SNPs in G1/S checkpoint regulators and their impact on the susceptibility to childhood leukemia. *Blood*. 2007 Jan 15;109(2):683-92.
10. Sinnett D, N'Diaye N, St-Onge P, **Healy J**. [Childhood leukemia: a genetic disease!] *Med Sci (Paris)*. 2007 Nov;23(11):968-74. French.
11. Marinou I, **Healy J**, Mewar D, Moore DJ, Dickson MC, Binks MH, Montgomery DS, Walters K, Wilson AG. Association of interleukin-6 and interleukin-10 genotypes with radiographic damage in rheumatoid arthritis is dependent on autoantibody status. *Arthritis Rheum*. 2007 Aug;56(8):2549-56.

SCIENTIFIC MEETINGS

1. 5th Annual Canadian Genetic Epidemiology & Statistical Genetics Meeting
King City, ON - May 2009
Bourgey M, **Healy J**, St-Onge P, Massé H, Sinnett D, Roy-Gagnon MH. A novel statistical strategy for the quantification of mating asymmetry in human populations.
2. 18th Annual meeting of the International Genetic Epidemiology Society
Kahuku, HI - October 2009
Bourgey M, **Healy J**, Roy-Gagnon MH, Sinnett D. Maternal-mediated genetic effects in early-onset diseases: an evaluation of methods.
3. 59th Annual meeting of the American Society of Human Genetics
Honolulu, HI - October 2009
Healy J, Bourgey M, Richer C, Larivière M, Dionne J, Roy-Gagnon MH, Sinnett D. The impact of maternal genetic effects in the initiation of childhood acute lymphoblastic leukemia.
4. 4th Annual Canadian Genetic Epidemiology & Statistical Genetics Meeting
Harrison Hot Springs, BC - May 2009
Healy J, Bourgey M, Roy-Gagnon MH, Sinnett D. The role of maternal genetic effects in the onset of childhood acute lymphoblastic leukemia.
5. Journées génétiques du Réseau de Médecine Génétique Appliquée
Québec, QC - May 2008

Healy J, Dionne J, Larivière M, Ouimet M, Gagné V, N'Daye N, Beaulieu P, Labuda D, Sinnett D. Regulatory genetics in cancer research – the childhood leukemia story.

6. The American Association for Cancer Research Special Conference on Approaches to complex pathways in molecular epidemiology. Albuquerque, NM - June 2007

Healy J, Bélanger H, Dionne J, Larivière M, Sinnett D. Impact of regulatory genetics in the susceptibility to childhood leukemia.

7. The Human Genome Organization's 12th Human Genome Meeting. Montréal, QC - May 2007

Healy J, Bélanger H, Larivière M, Labuda D, Sinnett D. Functional promoter SNPs in TGFB1 and the susceptibility to childhood leukemia.

8. The Annual meeting of the American Association for Cancer Research Los Angeles, CA - April 2007

Healy J, Bélanger H, Larivière M, Labuda D, Sinnett D. Functional promoter SNPs in TGFB1 and the susceptibility to childhood leukemia.

9. Genome Canada's International Conference Quebec, QC - October 2006

Healy J, Bélanger H, Beaulieu P, Larivière M, Labuda D, Sinnett D. Promoter SNPs in G1/S Checkpoint Regulators and their Impact on the Susceptibility to Childhood Leukemia.

10. European Mathematical Genetics Meeting Cardiff, Wales, UK - April 2006

11. The American Association for Cancer Research Special Conference for Cancer Research; Cell Cycle and Cancer: Pathways and Therapies Fort Lauderdale, FL - December 2004

Healy J, Bélanger H, Vasquez H, Moghrabi A, Sinnett D. Cyclin-dependent kinase inhibitors: regulatory polymorphisms and their impact on the susceptibility to childhood leukemia.

12. 54th Annual Meeting of the American Society of Human Genetics Toronto, ON - October 2004

Healy J, Bélanger H, Vasquez H, Moghrabi A, Sinnett D. Impact of Regulatory Polymorphisms in Genes Involved in Cell Cycle Control on the Susceptibility to Childhood Leukemia.

ADDITIONAL SCIENTIFIC TRAINING

2009	Scientific Communication Workshop Sainte-Justine Hospital Research Centre, Montreal, Qc
2009	Montreal Spring School of Population Genomics and Genetic Epidemiology Collège Jean-de-Brébeuf, Montreal, Qc
2005	Short Course in FBAT and PBAT Harvard School of Public Health, Boston, MA
2004	Summer Institute in Statistical Genetics North Carolina State University, Raleigh, NC
2002 – 2003	Summer Research Assistant in Molecular Medicine Ottawa Health Research Institute, Ottawa, ON

OTHER SCIENTIFIC ACTIVITIES

- Reviewer for PLoS ONE
- Member of the American Society of Human Genetics
- Member of the International Genetic Epidemiology Society
- Organizing committee: Montreal Spring School of Population Genomics and Genetic Epidemiology; May 2009, Collège Jean-de-Brébeuf, Montreal, QC
- Student supervision :
 - 3rd year Genomics and Proteomics Master's student (K. Benhamiza) from the Université des Sciences et Technologies de Lille, France; March-September 2007
 - 2nd year Biochemistry student (C. Malouf) from the University of Montreal Biochemistry Department; June-September 2006
- Participated in the Canadian Gene Cure Foundation's Gene Researcher for a Week mentorship program; March 2005 & April 2008

